

Journal of Applied Psychology

Edited by

Donald G. Paterson

University of Minnesota

7385

Consulting Editors

George K. Bennett, *Psychological Corporation*
Walter V. Bingham, *Washington, D. C.*
Harold E. Burtt, *Ohio State University*
Allen L. Edwards, *University of Washington*
Clifford E. Jurgensen, *Minneapolis Gas Co.*
Irving Lorge, *T. C. Columbia University*
Quinn McNemar, *Stanford University*

Alexander Mintz, *City College of New York*
James P. Porter, *Danville, Illinois*
Julian B. Rotter, *Ohio State University*
Edward K. Strong, Jr., *Stanford University*
Donald E. Super, *T. C. Columbia University*
Morris S. Viteles, *University of Pennsylvania*
Alfred C. Welch, *Knox-Reeves, Minneapolis*

Volume 34, 1950



Published Bi-monthly by the American Psychological Association, Inc.
Prince and Lemon Sts., Lancaster, Pa.

Entered as second-class matter, August 19, 1943, at the post office at Lancaster, Pa., under the Act of March 3, 1879

Acceptance for mailing at the special rate of postage provided for in paragraph (d-2), Section 34.40,
P. L. & R. of 1948, authorized October 10, 1947

Copyright, 1950, by The American Psychological Association, Inc.

Contents of Volume 34

Articles

Ammons, R. B., Butler, M. N. and Herzog, S. A. A Projective Test for Vocational Research and Guidance at the College Level.....	198
Baas, M. L. Kuder Interest Patterns of Psychologists.....	115
Barnette, W. L., Jr. The Non-Respondent Problem in Questionnaire Research...	397
Barnette, W. L., Jr. Reactions of Veterans to Counseling.....	399
Berdie, R. F. Scores on the Strong Vocational Interest Blank and the Kuder Preference Record in Relation to Self Ratings.....	42
Bittner, R. H. and Rundquist, E. A. The Rank-Comparison Rating Method.....	171
Brayfield, A. H. and Reed, P. A. How Readable are Occupational Information Booklets?.....	325
Browne, C. G. Study of Executive Leadership in Business. II. Social Group Patterns.....	12
Browne, C. G. Study of Executive Leadership in Business. III. Goal and Achievement Index.....	82
Browne, C. G. The Concentric Organization Chart.....	375
Burke, L. K. and Taylor, E. K. Rating Training and Experience.....	381
Campbell, D. T. and Mohr, P. J. The Effect of Ordinal Position upon Responses to Items in a Check List.....	62
Carr, E. R. and Rothe, H. F. Validity of an Objectivity Key on a Short Industrial Personality Questionnaire.....	178
Carter, H. L. J. A Combined Oral Reading and Psychogalvanic Response Technique for Investigating Certain Reading Abilities of College Students.....	267
Cook, Lt. Comdr. E. B. and Wherry, R. J. A Factor Analysis of MMPI and Aptitude Test Data.....	260
Corso, J. F. and Lewis, D. Preferred Rate and Extent of the Frequency Vibrato....	206
Cottle, W. C. Card Versus Booklet Forms of the MMPI.....	255
Cover, C. B. and Pressey, S. L. Age and Route Sales Efficiency.....	229
Davis, M. K. and Tiffin, J. Cross Validation of an Abbreviated Point Job Evaluation System.....	225
de Cillis, O. E. and Orbison, W. D. A Comparison of the Terman-Miles M-F Test and the Mf Scale of the MMPI.....	338
DuBois, P. H. and Watson, R. I. The Selection of Patrolmen.....	90
Dunlap, J. W. The Effect of Color in Direct Mail Advertising.....	280
Edwards, A. S. The Myth of Chronological Age.....	316
Ely, J. H., Kephart, N. C. and Tiffin, J. Ortho-Rater Norms and Sex Differences...	232
Engelhardt, O. E. de Cillis. The Minnesota Clerical Test: Sex Differences and Norms for College Groups.....	412
England, A. O. Getting Your Message Across by Plain Talk.....	182
Farr, J. N. Readability and Interest Values in an Employee Handbook.....	16
Flesch, R. Measuring the Level of Abstraction.....	384
Ford, A., Rigler, D. and Dugan, G. E. Point Centering of Signals on an Area.....	429
Ford, A. H. Prediction of Academic Success in Three Schools of Nursing.....	186
Frandsen, A. N. The Wechsler-Bellevue Intelligence Scale and High School Achievement.....	406
Gaudet, F. J., Carli, A. R. and Dennegar, L. S. Attitudes of Veterans toward Vocational Guidance Services.....	347

Gilbert, C. The Guilford-Zimmerman Temperament Survey and Certain Related Personality Tests	394
Golin, E. and Lysterly, S. B. The Galvanic Skin Response as a Test of Advertising Impact	440
Goodman, C. H. The MacQuarrie Test for Mechanical Ability. IV. Time and Motion Analysis	27
Gray, J. S. Custom Made Systems of Job Evaluation	378
Gray, J. S. and Prevetta, P. Fluorescent Light Versus Daylight	235
Griffith, J. W., Kerr, W. A. and Mayo, T. B., Jr. Changes in Subjective Fatigue and Readiness for Work During the Eight-Hour Shift	163
Hay, E. N. Cross-Validation of Clerical Aptitude Tests	153
Hay, E. N. The Application of Weber's Law to Job Evaluation Estimates	102
Hayes, P. M., Jenkins, J. J. and Walker, B. J. Reliability of the Flesch Readability Formulas	22
Jenkins, W. L., Maas, L. O. and Rigler, D. Influence of Friction in Making Settings on a Linear Scale	434
Johnson, R. H. and Bond, G. L. Reading Ease of Commonly Used Tests	319
Jones, F. N. and Smith, C. J. Visual Skill and Performance in a Meat Packing Plant	313
Jones, M. H. The Adequacy of Employee Selection Reports	219
Jurgensen, C. E. Intercorrelations in Merit Rating Traits	240
Jurgensen, C. E. Overall Job Success as a Basis for Employee Ratings	333
Kates, S. L. Rorschach Responses, Strong Blank Scales, and Job Satisfaction Among Policemen	249
Keating, E., Paterson, D. G. and Stone, C. H. Validity of Work Histories Obtained by Interview	6
Kephart, N. C. and Besnard, G. G. Visual Differentiation of Moving Objects	50
Kephart, N. C. and Mason, J. M. Acuity Differences between the Two Eyes and Job Performance	423
Kerr, W. A. Accident Proneness of Factory Departments	167
Klugman, S. F. Spread of Vocational Interests and General Adjustment Status	108
Layton, W. L. An IBM Card Profile for the Strong Vocational Interest Blank	415
Levine, A. S. Construction and Use of Verbal Analogy Items	105
Levine, A. S. Minnesota Psycho-Analogies Test	300
Link, H. C. and Freiberg, A. D. To What Extent Have the American People Accepted Socialism?	88
Lord, F., Cowles, J. T. and Cynamon, M. The Pre-Engineering Inventory as a Predictor of Success in Engineering Colleges	30
Mandell, M. M. The Administrative Judgment Test	145
Miller, H. K., Jr. An Exploratory Study of Linear Interpolation	367
Mitchell, M. B. and Rothe, H. F. Validity of an Emotional Key on a Short Industrial Personality Questionnaire	329
McCormick, E. J. An Analysis of Visual Requirements in Industry	54
Nahm, H. A Follow-Up Study on Satisfaction with Nursing	343
Navran, L. The Super-Roper Technique as a Measure of Interest in Nursing	417
Nesberg, L. S. and Smith, K. U. Measurement of a Complex Psychomotor Performance by Means of a Printed Test	309
Owens, W. A. An Aptitude Test for Veterinary Medicine	295
Owens, W. A., Jr. A Difficult New Test of Mechanical Comprehension	77
Page, H. E. A Note on Norms for the Purdue Industrial Mathematics Test and the Adaptability Test	306
Pashalian, S. and Crissy, W. J. E. How Readable Are Corporate Annual Reports?	244

Paterson, D. G. Report on the Journal of Applied Psychology for 1949.....	285
Peatman, J. G. and Hallonquist, T. Geographical Sampling in Testing the Appeal of Radio Broadcasts.....	270
Poruben, A., Jr. A Test Battery for Actuarial Clerks.....	159
Pronko, N. H. and Herman, D. T. Identification of Cola Beverages. IV. Postscript	68
Prothro, E. T. and Perry, H. T. Group Differences in Performance on the Meier Art Test.....	96
Ramond, C. K., Rachal, L. H. and Marks, M. R. Brand Discrimination among Cigarette Smokers.....	282
Romney, A. K. The Kuder Literary Scale as Related to Achievement in College English.....	40
Ross, S. A Study of Shooting Glasses by Means of Firing Accuracy.....	118
Rothe, H. F. Use of an Objectivity Key on a Short Industrial Personality Question- naire.....	98
Smith, A. J. Menstruation and Industrial Efficiency. I. Absenteeism and Activ- ity Level.....	1
Smith, A. J. Menstruation and Industrial Efficiency. II. Quality and Quantity of Production.....	148
Stern, B. Upper versus Lower Case Copy as a Factor in Typesetting Speed for Linotype Trainees.....	351
Travers, R. M. W. and Wallace, W. L. Inconsistency in the Predictive Value of a Battery of Tests.....	237
Wagner, R. F. Critical Requirements for Dentists.....	190
Wallace, S. R., Jr. and Whitney, A. G. The Prediction of Persistency in Premium Payment.....	131
Weitz, J. Verbal and Pictorial Questionnaires in Market Research.....	363
Wesley, S. M., Corey, D. Q. and Stewart, B. M. The Intra-Individual Relationship between Interest and Ability.....	193
Wiener, D. N. and Simon, W. Personality Characteristics of Embalmer Trainees...	391
Williams, A. C., Jr. and Roscoe, S. N. Evaluation of Aircraft Instrument Displays for Use with the Omni-Directional Radio Range (VOR).....	123
Woods, W. A. and Boudreau, J. C. Design Complexity as a Determiner of Visual Attention Among Artists and Non-Artists.....	355

Book Reviews

Allport's The Individual and His Religion: Paul E. Meehl.....	451
Anastasi and Foley's Differential Psychology: Gladys C. Schwesinger.....	444
Bellows' Psychology of Personnel in Business and Industry: Albert S. Thompson.....	136
Bellows and Rush's Workbook in Personnel Methods: Albert S. Thompson.....	448
Bennett and Cruikshank's A Summary of Clerical Tests: Donald E. Super.....	214
Boynton's Selecting the New Employee: Robert N. McMurry.....	214
Brouwer's Student Personnel Services in General Education: Milton E. Hahn.....	72
Calhoon's Problems in Personnel Administration: Albert S. Thompson.....	444
Cavan, Burgess, Havighurst, and Goldhamer's Personal Adjustment in Old Age: Albert R. Chandler.....	217
Chapanis, Garner, and Morgan's Applied Experimental Psychology. Human Fac- tors in Engineering Design: Jack W. Dunlap.....	213
Drake and Drake's A Human Relations Casebook for Executives and Supervisors: Albert S. Thompson.....	287
Flesch's The Art of Readable Writing: James J. Jenkins.....	71
Gilbert and Weitz's Psychology for the Profession of Nursing: Helen Nahm.....	139

Hoppock's Group Guidance; Principles, Techniques, and Evaluation: Milton E. Hahn.....	290
Hovland, Lumsdaine, and Sheffield's Experiments on Mass Communication. Studies in Social Psychology in World War II, Vol. III: Allen L. Edwards.....	139
Libo's Attitude Prediction in Labor Relations—A Test of Understanding: Edwin E. Ghiselli.....	371
Massing's Rehearsal for Destruction: Harrison G. Gough.....	447
Mathewson's Guidance Policy and Practice: Donald E. Super.....	141
McCord and Witheridge's Odors: Physiology and Control: Josef Brozek.....	447
Mossin's Selling Performance and Contentment in Relation to School Background: Daniel Raylesberg.....	287
Pease's Machine Computation of Elementary Statistics: Kenneth E. Clark.....	213
Prentice-Hall, Inc. The New Cure for White Collar Unrest: Clifford E. Jurgensen....	138
Pressey's Educational Acceleration: Appraisals and Basic Problems: John W. Gustad	448
Reynolds and Shister's Job Horizons: C. E. Jurgensen.....	70
Schramm's Mass Communications: Alfred C. Welch.....	446
Stuit, Dickson, Jordan, and Schloerb's Predicting Success in Professional Schools: Barbara A. Kirk.....	445
Super's Appraising Vocational Fitness: J. R. Wittenborn.....	136
Tinkelman's Difficulty Prediction of Test Items: Charles I. Mosier.....	452
Van Der Lugt's V. D. L. Adult Psychomotor Test Series for the Measurement of Manual Ability and V. D. L. Psychomotor Test Series for Children for the Measurement of Manual Ability: Robert H. Seashore.....	371
Vernon and Parry's Personnel Selection in the British Forces: Dewey B. Stuit.....	288
Wagner's Selective Job Placement: A. A. Canfield.....	70
Ward's Out-of-School Vocational Guidance: Fred M. Fowler.....	450
Warner, Gardner, Henry, and Haggard's Identifying and Developing Potential Leaders: Ralph R. Canter, Jr.....	289
Weitzman and McNamara's Constructing Classroom Examinations—A Guide for Teachers: Walter W. Cook.....	216
Weston's Sight, Light and Efficiency: Miles A. Tinker.....	449
Williamson's Trends in Student Personnel Work: Albert S. Thompson.....	72

Miscellaneous

New Books, Monographs, and Pamphlets.....	74, 143, 218, 292, 373, 454
---	-----------------------------

Journal of Applied Psychology

VOL. 34, No. 1

FEBRUARY, 1950

Menstruation and Industrial Efficiency. I. Absenteeism and Activity Level *

Anthony J. Smith

University of Kansas

A survey of the experimental work directly related to the influence of menstruation upon industrial efficiency reveals that investigators have been primarily concerned with absenteeism, with much less emphasis placed upon production. Studies indirectly related to the problem have dealt with such things as subjective changes and performances of various types, such as learning, steadiness, satiation, etc.¹

Many of these reports are to be criticized on several grounds. Some of the studies yielded generalizations based on pitifully small numbers of individuals. Social status and age were often disregarded. The suitability of the subjects, with reference to physical condition, was often given less consideration than the availability of the subjects. Analyses often contrasted menstrual with nonmenstrual data, rather than comparing the finer components of the menstrual cycle. Fluctuations occurring during menstruation were sometimes overemphasized, whereas other equally great fluctuations were ignored. The factor of suggestion was occasionally uncontrolled. Preconceived personal opinions at times seem to have been more powerful in determining conclusions than the actual data obtained. Statistical treatment of the data was inadequate in some studies, almost completely missing in others, with the result that the basis for inference was absent or uncomfortably shaky.

In spite of these criticisms, some facts seem

to stand out. Menstrual absences, despite occasional conflicting reports, would appear to be, in general, rather low and of much less importance to an employer than many other factors contributing to absenteeism. The numerous subjective changes, however (depression, increased emotionality, anxiety, fatigue, irritability, cramps, headache, etc.), sometimes classified as "pre-menstrual tension," are reported by subjects with great frequency. It would not be illogical to hypothesize that these changes would be reflected in the efficiency of a woman worker.

Procedure and Analytic Techniques

In the present investigation, workers in the electrical department of an aircraft factory and in two separate garment companies were studied. The criteria of industrial efficiency were absence rate, activity level, quality of production and quantity of production. The first two will be discussed in this paper.

Aircraft Factory. Those women who served as subjects in the aircraft factory were working at various tasks such as assembling electrical equipment, soldering, and constructing wiring circuits. They were paid a straight wage with provisions for higher overtime rates. The full-time secretary of the group kept accurate records of all leaves of absence, transfers, absences and tardinesses, and these records were made available for analysis. In addition to the use of absence rate as a criterion of efficiency, it was decided to examine a characteristic of behavior often commented on but little studied, i.e., activity level. It has been widely assumed that a woman in her menstrual period exhibits a lowering in general activity, an idea probably stemming from an unsubstantiated belief that a considerable number of women are forced to retire from general activity during this "sick" period. The lowering in activity level is presumably of such dimensions as to be readily observable. If this were true, such a change should be appar-

* This material is derived from the author's dissertation submitted in April, 1945 in partial satisfaction of the requirements for the Ph.D. degree at the University of California, Los Angeles. The author is deeply indebted to Dr. Roy M. Dorcus for his guidance and criticism.

¹ For an excellent review of this material see the article by Georgene H. Seward (6).

ent to someone who has been specifically instructed to look for it. It was evident that the leadmen in the group were in most intimate contact with the women and would be best qualified to make judgments concerning their activity levels. Consequently, they were required to turn in *daily* ratings on a form requesting the checking at the end of the day of one of the following: (a) Very energetic and industrious; (b) Fairly industrious, quite active; (c) Works in a slow, leisurely manner; and (d) Relatively inactive, works very slowly. The leadmen, of course, were not acquainted with the purpose of the study.

Menstrual data were collected through the cooperation of the personnel and physical education departments. It was absolutely necessary that none of the women participating recognize the true nature of the experiment for the results might then more closely reflect the anticipated effects of menstruation than its actual effects. To accomplish this, the experiment was described as being concerned only with the duration of the menstrual cycle and of the menstrual phase of the cycle.

A group of thirty-eight women between the ages of 17 and 44 had been selected on the basis of age (an arbitrary limit of 45 being set) and the availability of daily inspection records. These women were then assembled on company time and the prepared statement of the "purpose" was read by a member of the personnel staff. It was emphasized that participation would be deeply appreciated but that no one should fail to assert herself should she feel disinclined to supply the information. All offered to cooperate. They were then informed that a daily contact would be made by a member of the physical education department to ascertain the phase of the menstrual cycle. The women were only required to inform the person whether or not they were menstruating.

Information on absences, ratings of activity level, and menstruation was gathered for a period of forty-one days, during August and September, 1943. By the end of this period the number of women about whom sufficient information had been gathered to permit an analysis had been reduced to twenty-nine.

For purposes of analysis, the menstrual cycle was broken down in two ways. First, a simple menstrual (period of flow)-nonmenstrual division was made. Second, the cycle was broken down into a five day premenstrual period, a menstrual (bleeding) period, a seven day postmenstrual period, and an intermenstrual period.

The effect of the menstrual cycle upon absence rate was examined by determining the number of absences and number of days in attendance for all women for each of the four phases of the cycle. These data were then recorded in a two-by-four table and the significance of the variations was tested by means of the chi-square test. The nonmenstrual days were then combined in one group and the menstrual-nonmenstrual analysis was performed.

A group analysis of activity level was made by analyzing the ratings of the leadmen, again using a chi-square technique. Because of the low frequency

with which rank four was assigned, it was necessary to combine ranks three and four.

Parachute Factory. The second study was carried out in a local garment factory in which parachutes were being made. A survey of the plant revealed that it provided an ideal set-up for such an investigation as this. An overwhelming majority of the employees were women. There was a great diversity in the types of work performed in the plant, permitting a wide sampling of tasks, and yet, except on rare occasions, the women continued to work at the same jobs. Moreover, three shifts were available for study. Finally, a piece rate system of payment was in use, with the result that excellent records were kept by the management.

Forty-six women were selected for investigation so that there would be the most equitable distribution of women and jobs with reference to shift, age of employee, mental difficulty of work, physical difficulty of work, and the necessity for standing or sitting continuously on the job. These women were contacted and agreed to cooperate.

In addition to the analysis of the entire group, separate analyses were also made for each of the following thirteen groupings: three shifts, three age groups, three levels of mental difficulty, two levels of physical difficulty, and standing *vs.* sitting jobs.

Garment Factory. The last part of this investigation was carried out in a garment factory and was made possible through the cooperation of the personnel department and the local labor union.

The union gathered data on daily earnings, presumably to obtain information about salaries that would permit of a more enlightened discussion at union-management arbitrations. These data were recorded on mimeographed forms covering weekly periods, from which daily attendance records could then be derived.

The workers were contacted by the author's wife and information concerning menstrual cycles was requested with the same apparent "purpose" being quoted. At no time was any connection made between menstrual records and earning records.

Absences were then analyzed in the manner previously described.

Results

Activity Level. Results from the investigation of activity level are presented in Table 1. The analyses of these data indicate that leadmen cannot detect any reliable change in rate of working during the menstrual cycle. When the four phases of the menstrual cycle are examined, it is found that somewhat lower ratings were given during the menstrual and premenstrual phases. However, the probability value corresponding to the obtained chi-square was .10. When the menstrual and nonmenstrual components were contrasted, the probability value was .40.

It would seem justifiable to conclude that if real differences in activity level among the

Table 1

Activity Level: Frequency of Assignment of Various Ratings During the Phases of the Menstrual Cycle

Rating	Pre-menstrual	Men-strual	Post-menstrual	Inter-menstrual
1	50 (56.92)*	56 (61.56)	73 (74.34)	130 (116.16)
2	66 (64.83)	70 (70.12)	95 (84.67)	121 (132.30)
3-4	31 (25.24)	33 (27.30)	24 (32.97)	49 (51.51)

* Values in parentheses refer to expected frequencies.

various phases of the menstrual cycle actually exist, they would appear to be of such small magnitude as to warrant little or no consideration in this industrial situation.

Absences. The criterion of efficiency most adequately studied during this investigation was the absence rate of the worker. A total of 91 women acted as subjects and absence records for approximately 3800 potential working days were examined (man-days). The data for the workers in each of the three factories are presented in Table 2. It should be emphasized that menstrual absences are merely those that have occurred during the menstrual period and could conceivably be attributed to dysmenorrhea, colds, injury, personal business, indifference, etc. If it had been possible to secure adequate data on causes of absences, it seems likely that the results would have been in close agreement with Gafaer's work, in which it was discovered that absences attributable to dysmenorrhea were responsible for an average loss of only .29 days per person per year (2).

Information gathered at the aircraft factory revealed that menstrual absences occurred less frequently than might have been expected on the basis of chance, while premenstrual absences were more frequent. However, this trend was not significant ($P > .30$).

The more exhaustive study of the total group at the parachute factory yielded low menstrual absence rates and high postmenstrual absence rates. In the sub-groups no significant differences were found in the menstrual-nonmenstrual analyses. However, in

the examination of the four component phases of the menstrual cycle, significance was encountered in three instances. The women who worked on the second shift showed relatively high postmenstrual and low premenstrual absence rates ($P = .05$). Women in the age range 29-38 also manifested high postmenstrual and low menstrual absence rates ($P = < .01$). Finally, the women in the age range 39-50 had high premenstrual and low menstrual absence rates ($P = .02$).

In the first two instances (second shift, ages 29-38) excessive absences do not occur during those parts of the menstrual cycle that have been presumed to be most often characterized by the presence of unpleasant symptoms or disabling effects. The reasons for the appearance of high postmenstrual absence rate in these two cases are unknown but it is possible to speculate that it might be a function of the following factors: a tendency toward decreased social activity during the menstrual period, a decrease in the frequency of sexual relations during the period of flow (4), an increase in the frequency of intercourse during the postmen-

Table 2
Attendance Records for Total Groups

	Pre-menstrual	Men-strual	Post-menstrual	Inter-menstrual
<i>Aircraft Factory</i>				
Absent	11 (7.89)*	5 (8.13)	12 (10.14)	14 (15.83)
Working	150 (153.11)	161 (157.87)	195 (196.86)	309 (307.17)
<i>Parachute Factory</i>				
Absent	21 (22.23)	24 (26.18)	40 (32.30)	66 (70.39)
Working	328 (326.77)	387 (384.82)	467 (474.70)	1039 (1034.61)
<i>Garment Factory</i>				
Absent	1 (5.29)	11 (6.43)	8 (5.80)	11 (13.55)
Working	83 (78.71)	91 (95.57)	84 (86.20)	204 (201.45)

* Values in parentheses refer to expected frequencies.

strual period (4), an increase in the strength of the sexual drive during the postmenstrual period (1, 3), and a greater proportion of men working on the first and third shifts combined than on the second shift.

For the final sub-group yielding significance (ages 39-50) it might seem at first glance that earlier conceptions are substantiated. However, if the intermenstruum is accepted as a "normal" basis for comparison, this is not the case. The low menstrual absence rate offsets the high premenstrual rate. While the earlier conception is contraindicated, it would seem to be advisable to examine the high premenstrual rate for this older group in greater detail.

At the garment factory results were encountered that were significant at the 5 percent level. Here menstrual absences were high, premenstrual absences low.

There are several differences that exist between the garment factory and the parachute and aircraft factories that might be responsible for the obtained differences in absence rates.

In the war industries there was the possibility that patriotic motives might play a part in determining both group attitudes toward absenteeism and the likelihood that the individual would consider that slight discomforts would justify staying home.

The fact that the loss of a day would eliminate overtime payments for that week might cause many ailments to be overlooked, particularly those known to be nonprogressive and of relatively short duration. In the garment factory a straight piece rate system was employed, so absence caused no disproportionate loss.

Wages in general were appreciably higher in the war plants so that many of the better workers had left the garment factory, and those that remained might have been less conscientious concerning attendance, particularly during a period of labor shortage.

Differences may have existed in the cultural backgrounds of the subjects. The author does not have sufficient information available to argue that the garment workers come from a lower socio-economic level or that they are more likely to believe that menstruation has disabling effects. But in the light of the work of Sowton and Myers (7) in which it was shown that the detrimental effects of menstruation

tended to be greater among subjects from the lower socio-economic group, it would seem to be worthwhile to carry out further, more detailed investigations.

Finally, facilities for overcoming dysmenorrhea were more adequate in the parachute and airplane factories. The parachute factory had a nurse in attendance who treated the employees and gave suggestions for the relief of any discomforts. Employees could come into the dispensary to rest and obtain help whenever necessary. In the aircraft factory this same service was available. Furthermore, groups of women met and were given special instructions in matters of hygiene and exercises that would be beneficial in relieving fatigue, menstrual discomforts, etc. This process might perhaps work both through direct physiological changes, and through suggestion.

It should also be recalled that should no real differences exist in absence rate; approximately five samples out of each hundred will yield significant differences. Whether or not such an explanation might be defensible here, it is, of course, impossible to say.

Summary

This phase of the investigation was concerned with the relationship between the components of the menstrual cycle and two measures of industrial efficiency: activity level and absence rate. The subjects were 96 women in three separate factories.

The data indicate that:

1. There is no discernible change in activity level related to menstrual function.
2. Significant differences in absence rates among the phases of the menstrual cycle are encountered in three subgroups at the parachute factory and at the garment factory. High rates of absence are not characteristically found in any one phase of the cycle.

Received April 19, 1949.

References

1. Davis, Katherine B. *Factors in the sex life of twenty-two hundred women*. New York: Harpers, 1929.
2. Gafafer, W. M. (Ed.) *Manual of industrial hygiene and medical service in war industries*. Philadelphia: W. B. Saunders Co., 1943.

3. Howell, W. H. *Textbook of physiology*. (14th Ed.) Philadelphia: W. B. Saunders Co., 1941.
4. McCance, R. A., Luff, M. C., and Widdowson, E. E. Physical and emotional periodicity in women. *J. Hyg., Camb.*, 1937, 37, 571-611.
5. Rider, P. R. *An introduction to modern statistical methods*. New York: John Wiley, 1939.
6. Seward, Georgene H. Psychological effects of the menstrual cycle on women workers. *Psychol. Bull.*, 1944, 41, 90-102.
7. Sowton, S. C. M., and Myers, C. S. (I) and Bedale, E. M. (II). Two contributions to the experimental study of the menstrual cycle. I. Its influence on mental and muscular efficiency. II. Its relation to general functional efficiency. *Rep. industr. Fatigue Res. Board*, No. 45, London, 1928.

Validity of Work Histories Obtained by Interview *

Elizabeth Keating, Donald G. Paterson, and C. Harold Stone

Industrial Relations Center, University of Minnesota

The work of the employment interviewer, the vocational counselor, and the research worker in the field of labor mobility is dependent in large measure upon the validity of work histories obtained by interview. In view of the urgency of the problem and in view of the extensive literature on the interview, it is surprising to discover that so little evidence on the problem of validity has been reported.

Review of Literature

Search of the literature reveals only one study dealing directly with the problem of validity of work histories obtained by interview and in which tabular data are presented. Creamer and Coulter (7) in reporting a survey of unemployed mill workers in Manchester, New Hampshire, gave data on validity by comparing work history interview information with employer records.

The fact that only one study of this specific problem giving quantitative evidence has been unearthed is most surprising in view of the voluminous literature on personnel administration, personnel psychology, industrial relations, and labor market mobility. Furthermore, when one stops to realize the millions of situations in which routine employment practice requires "recording of work history of applicants and checking with previous employers" it is amazing to find no reports concerning the validity of such work histories. Books on personnel management such as those by Tead and Metcalf (23), Shefferman (21), Scott, Clothier et al. (20), Baridon and Loomis (2), Knowles and Thompson (11), Walters (27), Yoder (28), Pigors and Myers (18) and Jucius

(10) have been searched in vain for cited evidence. Books on personnel psychology such as those by Link (13), Hollingworth (8), Viteles (26), Laird (12), Jenkins (9), Burt (5), Maier (14) and Tiffin (24) are likewise unrewarding so far as *evidence* on this specific problem is concerned. Emphasis in these books is placed on the unreliability and low validity of the interview as a selection and placement device. Jenkins (9) and Laird (12), however, do mention validity of work histories obtained by interview and in both instances the impression is given that validity is low, but again evidence is not cited.

The most valuable treatment of the interview in the literature is that by Bingham and Moore (3). They present an excellent discussion of the problem of validity but the specific evidence cited concerns only their own study showing a high degree of validity of verifiable aspects of interviews with employees concerning employment guarantee plans and their attitudes and opinions toward such plans.

The literature on labor market mobility and studies of the unemployed in the labor market is likewise singularly free from detailed *evidence* concerning validity of work histories. Bowley and Burnett-Hurst (4) recognized the importance of the problem, collected pertinent data but failed to present the details. They merely stated: "Wage statements were checked in a considerable number of cases by definite facts from the employers, and the results show that there was no evident bias in the direction of overstatement or understatement, though there were mistakes" (4, p. 174). Bancroft (1) skirts the fringes of the problem by reporting a check on the accuracy with which unemployed workers on relief provide interviewers with statements concerning relief grants and date of latest registration at public employment offices. Palmer (17) is concerned with the reliability and consistency of responses given in labor market inquiries but cites no evidence on validity determined by checking with rec-

* The original data were secured by the Employment Stabilization Research Institute of the University of Minnesota in 1940-41. Principal findings of this study of the unemployed are reported in D. Yoder, D. G. Paterson, H. G. Heneman, C. H. Stone, et al. *Local labor market research*. Minneapolis: University of Minnesota Press, 1948. The present report was prepared with the assistance of Dr. Dale Yoder, Director of Industrial Relations Center, and Dr. H. G. Heneman, Jr., Assistant Director.

ords of previous employers. Even the excellent studies by the Yale researchers as reported by Reynolds and Shister (19) and by Noland and Bakke (16) fail to cite any evidence in regard to the validity of the work histories obtained in their extensive series of field interviews.

Clague, Couper and Bakke (6) tested the accuracy of workers' statements about prior jobs against employer records but failed to present quantitative evidence. Instead, they state: "As a result of these tests we have come to the following conclusions. A vast majority of the workers answer as accurately as they are able to do so. Only a very small fraction of the schedules gave evidence of attempts to mislead the interviewers . . . The replies with respect to wage rates were on the whole excellent . . . On time intervals, however, he is much less sure of himself . . . On the whole the data as presented in this study are accurate, and if the necessary corrections could be made the results would not be materially changed" (6, 129).

Myers and Maclaurin (15) report that, in addition to information secured directly from company records, they interviewed a sample of 233 workers and checked their answers against company records. Again, detailed tabular evidence is not given but statements are made which seem to imply low validity of work history reports. For example, "More than a third did not mention jobs which they were known to have had . . . It was seldom that a job held as long as six months was not mentioned by the worker. For this reason, we are inclined to attribute these omissions to 'poor memory.' . . . There were errors in estimating the length of particular jobs, and workers tended to overestimate rather than underestimate the lengths of the jobs they had held in the preceding three years. . . . A frequent discrepancy was found between the size of the weekly paycheck which the worker claimed to have earned and the amount which his employment records showed he had actually earned. In those cases where a comparison was possible, more than five times as many workers overestimated their earnings as underestimated them. Usually the difference was small, but the tendency was nonetheless im-

portant. . . . One conclusion that stands out from this analysis of discrepancies is that workers' memories or statements cannot be relied upon for detailed factual information concerning their work experience. They are frequently unable to remember all the jobs they have had, the order in which they had them, the dates of employment, the length of their jobs, and their earnings."

Creamer and Coulter's report (7) was a WPA study of the accuracy of work histories given by unemployed textile mill workers, 90 per cent of whom were foreign born. Records of 227 persons interviewed in their homes were compared with employer's records for the period 1930-1934. Results were reported by four tenure groups:

1. Those reporting one job with continuous employment during the 5 year period studied.
2. Those reporting one job but with one or two periods of no employment.
3. Those reporting more than one job with some periods of no employment.
4. Those reporting no employment during the five year period.

Results showed that only 14.1 per cent of the 227 cases reported their duration of employment with "complete accuracy." More than two-thirds of the cases involved over-statement of the length of employment. The conclusion is drawn that "this evidence suggests that in any industry characterized by intermittent employment, a work history of the fairly recent past based on the memory of the worker will tend to minimize the degree of intermittency by understating the number of jobs and by overstating the total duration of employment by appreciable amounts" (5, p. 342).

It is difficult to evaluate the Creamer and Coulter study in a straightforward manner. It is based upon home interviews of 227 former employees of a textile mill. The interviews apparently took place in November 1936 and covered prior employment with the Amoskeag Textile Mill extending backward in time from December 1934. The qualifications and training of the interviewers are not stated. Furthermore, employment in this particular company was characterized by "intermittency" so that it is little wonder that former employees

who were largely semi-skilled and foreign born were hazy and inaccurate in recalling the number of jobs held in the company and the duration of unemployment during each of the five years preceding December 1934. A further difficulty is that the results are presented in statistical tables which provide only totals and averages. These are difficult if not impossible to understand, let alone interpret. While the general impression of the low validity of work history data *as given* is probably correct, one should keep in mind that no precise index of validity is available and that the data themselves are provided by interviewers and interviewees under conditions that would be expected to promote a maximum of memory error.

In summary, the reports by Bowley and Burnett-Hurst and by Clague, Couper and Bakke give an impression of high validity whereas the reports by Myers and Maclaurin and by Creamer and Coulter give an impression of low validity. It is obvious that the importance of the problem in social science research and in personnel practice is such as to warrant citation of detailed evidence in such a manner that no one is forced to rely on author's conclusions but can interpret the evidence directly. There is urgent need for this kind of evidence. It is the aim of the present report not only to give this kind of evidence but also to give it in a form that is clear and unambiguous.

Procedure in the Present Study

The research reported in this paper is based on data gathered in studies of the occupational competence of unemployed persons in St. Paul conducted by the Employment Stabilization Research Institute in 1940-1942 (29). A random sample of unemployed persons who registered for employment at the St. Paul office of the USES during the period September, 1941, through February, 1942, was studied.¹

Extensive personal and occupational histories were secured from the registrants by interviewers on the staff of the Research Insti-

tute.² Specific information was obtained concerning the nature of job duties, dates of employment and separation, wage rates, reasons for separation, names of previous employers, etc., for every job held by each registrant since his entry into the labor force. Interviews were conducted in a clinical situation as part of a comprehensive program of individual case analysis. Those interviewed were unaware that their statements were to be subjected to independent verification. However, each person's full cooperation was obtained by informing him that the results of the interviewing and testing would be used not only for research but more importantly for counseling to enable the person to obtain the type of job and/or training for which he appeared to be fitted.

Verification and amplification of interviewees' statements were obtained by personal interviews with employers in the Twin City area by an "employer contact man" on the staff of the ESRI.³ Information was also obtained by mail from firms outside the metropolitan area. Using Toops' "follow-up" technique (25), an 80 per cent return was obtained on mailed inquiries.

Verification of employees' statements against company records calls for recognition of the possibility of error in company records. However, employer contacts in the study were made with great care and wherever a possibility of inaccuracy was noted by the employer contact man, this fact was stated in the case history report. Data from employers thus noted as questionable were not used. In this study, consequently, discrepancies between the statements of former employees and the employer are assumed to be the result of inaccurate information provided by the former employee. However, errors on the part of the interviewers in securing and recording work histories may have been involved.

Complete case histories for 385 unemployed

² These interviewers were first-year graduate students specializing in personnel psychology who had had one or more courses dealing at least in part with the general principles of interviewing. Before actual interviewing in the employment office began, each was given a minimum of one day of detailed instruction and supervised practice in conducting the type of interviewing required in this study. Close supervision was continued throughout the investigation.

³ Henry Morgan of the Research Staff was assigned the responsibility of securing employer reports.

¹ For further detailed discussion of the sample, see 29, p. 129 ff.

were available.⁴ Case histories for 236 registrants (157 men, 79 women) contained sufficient information to permit their inclusion in the present study. Employer reports on 373 jobs held provided information in one or more of the areas studied i.e., duration of job, weekly wage received, and nature of duties performed.

Because memory distortion would be expected to operate differentially in terms of the time these 373 previous jobs were held, it was necessary to classify each by a time interval category. Thus, a 0-12 months category included all jobs terminated within one year prior to the interview. The 13-24 months category included all jobs terminated from 13 months to 2 years prior to the interview, and so on. This breakdown was designed to reveal differences in validity of the occupational history between the immediate and more distant past.

Case histories of men and women were kept separate to facilitate study of sex differences in validity of job reports. Scatter diagrams for each time-interval were plotted to show the relation of the employer contact report to the interview record with respect to: 1. Weekly wages (male and female); and 2. Duration of job (male and female).

Results for One Year Time Interval

Wages. Table 1 shows the relationship between reported weekly wages and verified weekly wages for male workers on jobs held within one year of time of interview. Table 2 presents the same data for the women workers. Validity coefficients of +.90 and +.93 (based on ungrouped data) may be regarded as justifying the acceptance of wage data reports obtained under the circumstances surrounding this investigation with considerable faith. Furthermore, neither men nor women showed any consistent tendency to over-state or to under-state their wages.

Length (Duration) of Job. Distribution of data for jobs on which information was available on job duration is given in Tables 3 and 4. Again, the agreement between worker reports and employer records is close as revealed by validity coefficients of +.98.

⁴ These cases range from "professional" to "unskilled." See *Local labor market research* (29, 140-141) for distribution according to primary job classifications.

Table 1

Relationship between Reported Weekly Wages and Verified Weekly Wages for 116 Male Workers on Jobs Held 0-12 Months Prior to Interview (1940-1941)

Note: Pearson r for ungrouped data = +.90.

Reported Weekly Wages	Verified Weekly Wages							Total
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	
60-69							1	1
50-59					1	—		1
40-49				2	4	1		7
30-39			7	11	1			19
20-29		4	37	8			1	50
10-19	2	26	4					32
0-9	4	1						5
Total	6	31	48	21	6	1	2	115*

* One case not shown: reported weekly wage, \$105; verified weekly wage, \$100.

Table 2

Relationship between Reported Weekly Wages and Verified Weekly Wages for 61 Female Workers on Jobs Held 0-12 Months Prior to Interview (1940-1941)

Note: Pearson r for ungrouped data = +.93.

Reported Weekly Wage	Verified Weekly Wage						Total
	0-4	5-9	10-14	15-19	20-24	25-29	
25-29						2	2
20-24					3		3
15-19			4	19			24
10-14			24	3			27
5-9		3	1				4
0-4	2						2
Total	2	3	29	22	3	2	61

Duties of Job. In the 0-12 months time interval category, there were 144 male cases which contained data sufficient for study. Agreement between the interviewee's statement and the employer report of job duties was found in 94 per cent of the cases. There was a 96 per cent agreement for the 68 women. The nature of the data are such as to preclude the presentation of "scatter tables" or the computation of correlation coefficients.

In the small percentage of cases where the duties of the job as stated by the interviewee

Table 3

Relationship between Reported and Verified Duration of Employment in Months for 127 Male Workers on Jobs Held 0-12 Months Prior to Interview

Note: Pearson r for ungrouped data = $+.98$.

Reported Duration (Months)	Verified Duration (Months)						Total
	0-19	20-39	40-59	60-79	80-99	100-119	
80-99					2	1	3
60-79			1	6			7
40-59			5	2			7
20-39		14	2				16
0-19	93						93
Total	93	14	8	8	2	1	126*

* One case not shown: reported duration, 192 months; verified duration, 192 months.

Table 4

Relationship between Reported and Verified Duration of Employment in Months for 64 Female Workers on Jobs Held 0-12 Months Prior to Interview

Note: Pearson r for ungrouped data = $+.98$.

Reported Duration (Months)	Verified Duration (Months)						Total
	0-19	20-39	40-59	60-79	80-99	100-119	
100-119						1	1
80-99					1		1
60-79				1			1
40-59		1	1	1			3
20-39		3					3
0-19	53	1	1				55
Total	53	5	2	2	1	1	64

did not agree with the employer contact reports, there was a tendency to inflate the level of skill and responsibility in their jobs.

Results for the Longer Time Intervals

Because the number of cases involved in reports for time intervals greater than one year is small the results are presented in terms of correlation coefficients only.

Wages. Table 5 indicates that the correlations are positive and high for both men and women workers regardless of time interval. In the case of the men, the time interval extends

to six years. Of course, the numbers of cases is so small that the correlation coefficients must be accepted only with great reservations. The most that can be said is that the evidence does not indicate any definite drop in validity with the passage of time.

Duration of Employment. Table 6 also indicates that duration of employment is reported with a surprisingly high degree of validity even though time intervals from one to six years are involved. Of course, the numbers of cases involved are small and the correlations are not to be accepted at face value. Nevertheless, there is no evidence of decreasing validity with lapse of time.

Duties of Job. Data on reports of duties of the job, held at time intervals prior to interview greater than one year, were found to be surprisingly accurate. It is only the occasional applicant for employment who distorts his

Table 5

Correlations between Reported Weekly Wages and Verified Weekly Wages for Male and Female Workers on Jobs Held at Time Intervals Greater Than One Year

Time Interval (Job Termination Date Prior to Interview)	Men		Women	
	Number	r	Number	r
13-24 Months	24	.95	11	.97
25-36 Months	10	.94	8	.99
37-48 Months	19	.62	6	.91
49-60 Months	5	.99	—	—
61-72 Months	8	.76	—	—

Table 6

Correlations between Reported and Verified Duration of Employment for Male and Female Workers on Jobs Held at Time Intervals Greater than One Year

Time Interval (Job Termination Date Prior to Interview)	Men		Women	
	Number	r	Number	r
13-24 Months	33	.96	12	.99
25-36 Months	16	.99	7	.99
37-48 Months	22	.94	7	.94
49-60 Months	12	.97	—	—
61-72 Months	7	.88	—	—

previous work history under the conditions surrounding the present investigation.

Summary

In the extensive literature on the interview surprisingly little evidence exists with respect to the validity of work histories obtained by interview. Only one study was found bearing directly on the problem and this study was seriously defective.

The present study of applicants for employment in the St. Paul Office of USES was conducted on a research basis in a setting in which occupational guidance was stressed. Presumably little incentive existed for these applicants to distort their work histories. The validity of the work histories when checked by employer reports was found to be surprisingly high with respect to weekly wages, duration of employment and job duties. Validity remained high for histories secured for jobs held up to six years prior to the interview as well as for jobs held just prior to the interview. In terms of correlation coefficients, the validities may be generalized as being from $+ .90$ to $+ .98$.

Received October 17, 1949.

Early publication.

References

1. Bancroft, G. Consistency of information from records and interviews. *J. Amer. statist. Ass.*, 1940, 35, 377-381.
2. Baridon, Felix E., and Loomis, E. H. *Personnel problems*. New York: McGraw-Hill Book Co., 1931. Pp. 25-29.
3. Bingham, W. V., and Moore, B. V. *How to interview*. New York: Harper and Brothers, 1934 (rev. ed.). Ch. 12, pp. 190-216. (The 1942 revision omits the report of their validity study.)
4. Bowley, A. L., and Burnett-Hurst, A. R. *Livelihood and poverty*. London, England: G. Bell and Sons, Ltd., 1915. Pp. 222.
5. Burt, H. E. *Principles of employment psychology*. New York: Harper and Brothers, Rev. Ed., 1942. Pp. 405-475.
6. Clague, E., Couper, W. J., and Bakke, E. W. *After the shutdown*. New Haven: Institute of Human Relations, Yale University, 1934.
7. Creamer, D., and Coulter, C. W. *Labor and the shut-down of the Amoskeag Textile Mills*. Philadelphia: Work Projects Administration, National Research Project Report No. L-5, November, 1939. Pp. 342.
8. Hollingworth, H. L. *Judging human character*. New York: D. Appleton and Co., 1923. Pp. 263.
9. Jenkins, J. G. *Psychology in business and industry*. New York: John Wiley and Sons, Inc., 1945. Pp. 72-78.
10. Jucius, Michael J. *Personnel management*. Chicago: Richard D. Irwin, Inc., 1947. Pp. 175-200.
11. Knowles, A. S., and Thompson, R. D. *Industrial management*. New York: The Macmillan Co., 1944. Pp. 334-336.
12. Laird, D. A. *The psychology of selecting employees*. New York: McGraw-Hill Book Co., Inc., 1937 (3rd ed.). Pp. 99-119.
13. Link, H. C. *Employment psychology*. New York: The Macmillan Co., 1921. Pp. 435.
14. Maier, N. R. F. *Psychology in industry*. Cambridge: The Riverside Press, 1946. Pp. 426.
15. Myers, C. A., and Maclaurin, W. R. *The movement of factory workers*. New York: John Wiley and Sons, Inc., 1943. Pp. 111.
16. Noland, E. W., and Bakke, E. W. *Workers wanted*. New York: Harper and Brothers, 1949. Pp. 224.
17. Palmer, G. L. *The reliability of response in labor market inquiries*. Washington, D. C.: Technical Paper No. 22, Executive Office of the President, Bureau of the Budget, Division of Statistical Standards, July, 1942.
18. Pigors, P., and Myers, C. A. *Personnel administration*. New York: McGraw-Hill Book Co., Inc., 1947. Pp. 497.
19. Reynolds, L. G., and Shister, J. *Job horizons*. New York: Harper and Brothers, 1949. Pp. 102.
20. Scott, W. D., Clothier, R. C., et al. *Personnel management*. New York: McGraw-Hill Book Co., Inc., 1941.
21. Shefferman, N. W. *Employment methods*. New York: The Ronald Press Co., 1920.
22. Stead, W. H., and Shartle, C. L. *Occupational counseling techniques*. New York: American Book Co., 1940. Pp. 212.
23. Tead, O., and Metcalf, H. C. *Personnel administration*. New York: McGraw-Hill Book Co., Inc., 1933 (3rd ed.).
24. Tiffin, J. *Industrial psychology*. New York: Prentice-Hall, Inc., 1947 (2nd ed.).
25. Toops, H. A. The returns from follow-up letters to questionnaires. *J. appl. Psychol.*, 1926, 10, 92-101.
26. Viteles, M. S. *Industrial psychology*. New York: W. W. Norton and Co., Inc., 1932.
27. Walters, J. E. *Personnel relations*. New York: The Ronald Press Co., 1945.
28. Yoder, D. *Personnel management and industrial relations*. New York: Prentice-Hall, Inc., 1948 (3rd ed.). Pp. 894.
29. Yoder, D., and Paterson, D. G., et al. *Local labor market research*. Minneapolis: University of Minnesota Press, 1948.

Study of Executive Leadership in Business.

II. Social Group Patterns

C. G. Browne

Wayne University

This is the second in a series of articles presenting the following methods for the study of executive relationships and leadership in business: RAD index,¹ social group patterns and organizational contacts, sociometric pattern, and Goal and Achievement index.

The total study proceeded on the following hypotheses: (1) leadership is a process based upon the inter-relationships of individuals in a group that is working toward a goal which has been accepted as desirable; (2) executive function and leadership in business are processes of the interaction of social and working relationships within and outside of the executive groups; (3) executive and leader relationships can be analyzed through the application of methods which are not designed to measure personal executive traits as psychological entities.

The subjects for the study were 24 executives in a tire and rubber manufacturing company, representing all but one of the executives on the first, second, third, and fourth echelons of the business. They were classified into the following departmental groups: general administration, 4 cases; sales, 6; finance, 4; manufacturing, 8; and personnel, 2.

Social Group Patterns

As part of a longer interview, each executive indicated his first, second, and third choices of the following social groups in terms of their importance in his social life: group 1, individuals within the company, called the Company group; group 2, individuals outside of the company but with whom there was a business affiliation with the executive's own work and called the Outside and Business group; and group 3, individuals outside of the company with whom there was no business affiliation, called the Outside Only group.

Each executive also indicated the amount of social activity which he had with each social group using a three-point scale, the intervals designated as "large amount," "some," and "no" social life.

In calculating the score on social group patterns for each executive, the first social group choice was given six points, that is, the social group which was most important in the social life of the executive; second choice was given four points; and third choice, two points. For the amount of activity with each social group, "large" was assigned four points; "some," two points; and "no," zero points. The score for each executive was the product of the points allowed for the order of each social group choice and the amount of social activity with each group.

In Table 1 the social group pattern scores are tabulated by total and departmental groupings of executives.² For the total executive group and the sales, finance, and manufacturing departments, the three social groups ranked Outside Only, Company, Outside and Business. However, the general administration group ranked the Company social group first, followed in order by Outside and Business and Outside Only. Whereas Outside and Business was the second social group with the general administration executives, it was the third or least important group with all of the other executive groups and the total group. However, Outside and Business was a relatively more important social group with the sales executives than it was with finance or manufacturing. The finance executives indicated no Outside and Business social contacts, but their score for Outside Only social contacts was higher than for any other executive group.

In their total social activities, the general administration executives had the highest standing, followed in order by sales, manu-

¹ Browne, C. G. Study of executive leadership in business. I. The RAD index. *J. appl. Psychol.*, 1949, 33, 521-526.

² The personnel executives were included in the manufacturing group.

Table 1

Social Group Pattern Scores Tabulated by Total and Departmental Groups of Executives

Social Group	Executive Department	N	Mean Score	Median Score	Total Points
Company	Total	24	11.2	8.5	268
	GA*	4	18.0	20.0	72
	S	6	10.0	8.0	60
	F	4	8.0	8.0	32
	M	10	10.4	8.6	104
Outside and Business	Total	24	5.2	0.9	124
	GA	4	11.0	10.0	44
	S	6	9.3	4.0	56
	F	4	0.0	0.0	0
	M	10	1.6	0.8	16
Outside Only	Total	24	13.5	12.0	324
	GA	4	8.0	8.0	32
	S	6	13.3	16.0	80
	F	4	15.0	12.3	60
	M	10	13.6	11.7	136

* Code: GA—General Administration; S—Sales; F—Finance; M—Manufacturing.

facturing, and finance. Both general administration and sales were above the mean and the median for the total group, while finance and manufacturing were below.

From this analysis it appeared that a general social pattern developed which was based upon the type of work which the executive did or the department to which he was assigned. This pattern centered around the relative importance of the company or business contacts in social activities, the increase in their relative importance being observed in the general administration and sales departments. On the other hand, there was a complete absence of social business contacts in the finance department. A consideration of the individual executive's duties and work performed revealed that the social group choices followed the lines of necessity to a large degree in that certain types of social activity became necessary and essential to the performance of the individual as well as the departmental functions.

In Table 2, the social group pattern scores are tabulated by echelon level in the business.

The president and general manager, who constituted echelon one, had the highest possible score for each of the three social groups, indicating a large amount of social life with each social group. For him, the groups ranked as follows: Outside and Business, Company, Outside Only.

Echelon two was second in the total amount of social activity score, with the social groups ranked Company, Outside Only, Outside and Business. Echelon four preceded three in the total amount of social activity, this being strongly influenced by the scores of four sales executives who were in the fourth echelon. It was observed in the previous section that sales executives tended to have more social life. Both echelons three and four ranked the social groups Outside Only, Company, Outside and Business.

Thus, the executives at higher echelons exhibited a closer identification with the company through their social activities, the Company and Outside and Business social groups being more important than the Outside Only group. When this is combined with the departmental pattern of social choices, it appears that executives in general administration and sales and those in the higher echelons throughout the business had more social activity, particularly with those social groups which had a relationship with company operations.

Table 2

Social Group Pattern Scores Tabulated by Executive Echelon Level

Social Group	Echelon Level	N	Mean Score	Median Score	Total Points
Company	One	1	16.0	16.0	16
	Two	7	14.3	8.7	100
	Three	10	9.2	8.3	92
	Four	6	10.0	8.3	60
Outside and Business	One	1	24.0	24.0	24
	Two	7	7.0	3.5	48
	Three	10	2.0	0.6	20
	Four	6	3.3	3.7	20
Outside Only	One	1	8.0	8.0	8
	Two	7	10.3	11.3	72
	Three	10	15.2	12.3	152
	Four	6	15.3	14.0	92

Organizational Contacts

To study organizational contacts, each executive named all of the professional and social organizations to which he belonged, and indicated which memberships, if any, were paid by the company. Table 3 presents these results by executive department and echelon. The general administration group indicated the greatest number of social organization memberships per executive, followed by personnel, sales, finance, and manufacturing. It should be remembered, however, that the personnel group consisted of only two cases. However, the company was paying for social memberships for certain executives in the general administration and sales areas only. A study of the duties of these executives revealed that their activities with social organizations constituted an important aid to their company functioning. Obviously, then, the company was willing to carry the expense of at least some of these organizations.

The personnel group ranked highest in the number of professional organizations per executive, followed by general administration, finance, manufacturing, and sales. Of the six executives in the sales group, only the vice-president-sales and the sales manager belonged to any professional organizations. Each was a member of one organization, but these mem-

berships were not paid by the company. In finance, both the treasurer and the comptroller belonged to two professional organizations, and in each case, both memberships were paid by the company. The chief cost accountant belonged to three professional organizations, none of which were paid by the company. In the manufacturing group, although 12 professional memberships were held, only one of them, for the chief chemist, was paid by the company.

When the organizational memberships, and memberships paid by the company were analyzed on the basis of executive echelon groups, Table 3 shows a decrease in each variable from echelon one through echelon four. That is, executives in successively lower echelons belonged to fewer social organizations, fewer professional organizations, and fewer memberships were paid by the company for each type of membership. It might be said that as an executive advanced in this company into higher echelons, it could have been expected that he would become a member of more organizations, and that his work would be such that the company would be willing to pay for more of these memberships. On the other hand, it may have been that the organizational activities of these executives were characteristic of the individual executive operating in his own particular pattern and were not a particular function of his position in the business.

Table 3

Membership in Social and Professional Organizations Tabulated by Executive Department and Echelon

Membership in Social and Professional Organizations Tabulated by Executive Department and Echelon							
		Social Memberships (mean)		Professional Memberships (mean)		Total Memberships (mean)	
	N	Number	Paid by Co.	Number	Paid by Co.	Number	Paid by Co.
Executive Departmental Groups							
Total group	24	2.46	.46	1.67	.67	4.13	1.13
Genl. Adm.	4	4.50	1.75	2.75	1.75	6.25	3.50
Sales	6	2.50	.67	.33	0.00	2.83	.67
Finance	4	1.75	0.00	1.75	1.00	3.50	1.00
Manufacturing	8	1.50	0.00	1.50	.12	3.00	.12
Personnel	2	3.50	0.00	4.00	2.00	7.50	2.00
Executive Echelon Groups							
Echelon 1	1	10.00	5.00	4.00	4.00	14.00	9.00
Echelon 2	7	2.57	.86	1.86	1.00	4.43	1.86
Echelon 3	10	2.10	0.00	1.80	.40	3.90	.40
Echelon 4	6	1.67	0.00	.67	.17	2.34	.17

Conclusions

From this study of the social group patterns and the organizational contacts of a population of 24 executives in a tire and rubber manufacturing company, the following hypotheses may be advanced:

1. The social group choices and the amount of social activity of a business executive are in part determined by: (1) the work which he does within the business; (2) the department in which he works; (3) the echelon level into which he is classified within the organization.

2. Among the characteristics of executive performance and leadership in business are membership in social and professional organizations and memberships paid by the company, the number of all of these variables increasing as the executive advances in echelon level.

The question of whether the variables studied were a function of the job or the indi-

vidual in the job can be answered for business leadership in general only with horizontal studies of many executives in a wide variety of companies or with longitudinal studies over an extended period of time with specific companies or with a given group of executives.

For the present, however, the discussion presented here may prove to be of value in the selection and training of business executives since it suggests that (1) working relationships at higher echelons and in certain departments may extend beyond the confines of the office and the factory; (2) the individual who is aspiring to higher echelons or to certain departments may need to accept the probability that his executive duties for the company will extend to his social life; and (3) at least part of an executive's social life will be planned for the benefit which it will yield to the company and his own success in it.

Received May 17, 1949.

Readability and Interest Values in an Employee Handbook *

James N. Farr

University of Minnesota

Adequate communication between management and workers is essential to the success of modern personnel programs. The basis of such programs is an informed work force; a work force informed about company policies, plans, and competitive position; a work force with information as to the rules governing the organization of which they are a part, and an explanation of the "why" of these rules.

The means developed for this sharing of information include employee handbooks, house organs, company newspapers, bulletin boards, etc. These media are important, and much time and money are spent in their production. There is a growing literature dealing with the functions of these media, and the problems peculiar to each. There are available detailed discussions of how to prepare such media for publication, dealing with every step from formulation of editorial policy to distribution of the finished product.

Such discussions, adequate though they are for the problems discussed, do not go far enough. Distribution of the finished handbook or house organ is not the final step of the process. The process is not complete until the employees read what has been written *and understand what they read!*

Two aspects of the publication are important in this final step of the process; the attention and interest value of the publication, and the readability of the material in the publication. The fact that employee publications generally contain pictures and illustrations is an indication that the attention and interest problem has been recognized and is being met in part,—but only in part. It can never be wholly met so long as the second aspect—readability—is neglected, for a reader will not maintain interest in material that is difficult for him to understand.

* For an example of an excellent employee handbook the reader is referred to a handbook for Northwest Airlines employees entitled "Hello! I'm Topper." It was written for Northwest Airlines, Inc. by Mr. Wendell Knowles, and was the source for a number of the ideas used in the handbook discussed in this article.

Review of Literature

That the problem of readability is being neglected is evident. It is scarcely touched upon in the literature dealing with employee publications. Heron (5) who gives us an excellent discussion of the purposes, uses, and problems of communications media tells us only that the material must be readable and understandable. The National Foremen's institute (6) in its manual on how to prepare an employee handbook discusses in great detail the content and technical problems to be met in the publication. The problem of readability, however, is not discussed except for a statement that "two-dollar" words should be avoided, and that the writer should strive for simplicity. Biklen and Breth (2) omit discussion of the readability problem in their excellent book on employee publications. Bentley (1) in his book which deals in great detail with problems of editing an employee publication, dismisses the readability problem with the recommendation that newspaper style writing be used. This presumably means simplicity, familiar words, short sentences, and not too many adjectives and adverbs.

In view of the fact that the achievement of readable material in employee publications is essential if it is to have any value at all, the problem deserves much more careful consideration than this. What good is done by a publication which is perfect in its technical aspects and content, but which cannot be read with understanding by a considerable portion of the intended audience?

Rarely is any information given as to how to achieve readability. Apparently it is assumed that the writer of a handbook or house organ need only be aware of the fact that his material should be readable and that he will then see to it that it is so. Nothing is further from the truth. Seeing to it that a handbook or house organ is readable and will be read requires just as much effort and thought as

seeing to it that it includes the correct content and expresses the desired point of view. This fact is often overlooked, or not realized at all. Any personnel man who writes a handbook knows intuitively that he writes in a readable style. After all, it seems perfectly clear to him, and his colleagues all seem to understand what he has written. This is true, but the fact of the matter is that the handbook is not being written for his colleagues, but for the main body of the firm's employees. These people are not college graduates—a good many of them are not even eighth grade graduates—and material written by a college graduate, and understood by his colleagues, is not so readily understood by these workers. In fact, it may seem so difficult to them that they do not even attempt to read it.

This fact, that information published for workers must be written in language they can understand, if it is to hold their interest, seems to be self evident. Few would argue that this is not necessary. Most, in fact, would argue that they consider this factor carefully when publishing a handbook, for instance. This is no doubt true. Every effort may be made to make the handbook readable,—but for whom? The difficulty here seems to be in overlooking the fact that what is readable for one group of people is not necessarily readable for another group. The problem is to see that the handbook is readable for the intended reader, not for the writer.

This has been a difficult problem. It is easy to look at a piece of writing and tell whether it is hard to read or easy to read, in terms of your own scale of difficulty. It is much harder to judge the difficulty level in terms of the reading ability of someone else, such as a workman who has had an eighth grade education. Fortunately we now have several objective methods of estimating the reading difficulty level of written material. Application of the readability formulae developed by Flesch (3, 4, 4a), for example, gives a good estimate of reading difficulty, and of the educational level of those who will be able to read the material with understanding.

If the application of the Flesch formulae indicates that the material is written at a difficulty level too high for the intended audience, the chief problem then arises: how to

write at a level below that on which you are accustomed to write without seeming to "write down." Flesch (3, 4a) has made this problem much easier to solve. By following the rules he has laid down it is possible with practice to produce material which is both readable and well written.

Application of the Flesch formulae to management communications indicates that very often the material is written on a difficulty level far beyond the comprehension of the group to whom it is directed. Paterson and Jenkins (7) in their analysis of an information sheet for potential applicants in one large firm found that it was far too difficult for that group. The writer found in an analysis of 25 company house organs, and 25 union papers, that they were written at a level that effectively shut out a large portion of the intended readers.

This is obviously a ridiculous situation. It is futile to expend large amounts of time, energy, and money to publish a handbook or house organ which will not, or cannot be read with understanding by many of those for whom it is prepared.

Proposed Employee Handbook

Recently the writer was asked to analyze and revise a proposed employee handbook for a textile firm. The results of the analysis, and the steps taken in the revision to ensure a readable handbook will be discussed in the following pages.

Before beginning the revision, Flesch's readability formulae were applied to samples drawn from the proposed handbook. This sampling indicated that the difficulty level of the writing was far too high for the group of employees to whom it was to be given.

The first step, therefore, was to rewrite the handbook in much simpler language. This was done by following the rules given by Flesch (3, 4a). Long sentences were broken up into short simple sentences; words of one or two syllables were substituted for longer words wherever possible; statements were addressed directly to the reader wherever possible; and the material was personalized by the use of personal pronouns.

This resulted in a handbook which was much easier to read, but which was far too

long. It was now in a form which the workers *could* read but not in a form which they likely *would* read. For even though a handbook is written at a level the worker can understand, it will not hold his interest if it consists of page after page of words. It must be remembered that though the facts presented in a handbook are of value to a worker, they may not seem to him to be of immediate value. He will take them if they are easy to come by, but he will not go out of his way to sift these facts from paragraphs of useless words.

For this reason, a handbook should contain as few words as you can get by with rather than all that you can crowd into it. Get rid of words wherever possible. Put in a picture instead. Or leave white space. If nothing else, this will make the important facts and ideas stand out so that some of them get across even in a cursory reading.

The second step, then, was to strip away all excess words which served only to hide the ideas that were to be presented. The result of this step was to reduce the number of words from approximately 10,400 to approximately 2,800, and the number of pages from 39 to 28.

The revision now contained 28 pages with an average of 100 words per page. This meant there was ample white space, and ample room for pictures or illustrations. In the fourth step of the revision, illustrations were added.

The illustrations used were simple line figures. They serve to catch attention and to emphasize the rules or ideas being discussed. They also offer a painless method of "laying down the law." Consider the example in Figure 1.



Fig. 1. Example of line drawing to illustrate the company rule regarding excessive tardiness.

This is much less harsh than a statement to the effect that "if you are late too often, disciplinary action will be taken."

The revision was now complete. The handbook had been rewritten in simple language, excess words had been discarded, leaving the principal ideas emphasized on uncrowded pages, and simple illustrations had been added.

Below is reproduced a page from the original handbook, followed by the page from the revised handbook which dealt with the same subject matter.

Incentive Pay

The Company has installed wherever possible an incentive or piece work plan as a means of giving additional pay to those producing more than the average for a normal day's work. It is important that you, as an employee of the company, be fully familiar with the operation of this plan.

A trained time study person observes the operation being performed by an operator. The job is broken down into the work elements of which it is composed and a time study person is trained to recognize unusually good or below average effort and ability. If the time required to perform an operation is less than normal expectancy, the employee benefits from the extra effort. If an employee is working slower than normal, the reverse would be true. Provision is made for normal delays under personal and machine allowances.

Time standards are guaranteed against change as long as a method of operation and conditions surrounding the operation remain unchanged. However, in those instances where either the operator or the company feels that a rate is too low or too high, a re-timing may be requested and a change in rate negotiated. Any complaint on rates should be brought to the attention of your department supervisor. Time studies will be made in all instances where the department supervisor and the person in charge of time study activities feel that a re-timing should be made.

The management encourages any questions or suggestions concerning the operation of our incentive program. We are always ready to accept constructive comments on this subject because in the final analysis it will be mutually advantageous.

Compare the two sample pages shown. Can there be any doubt as to which can be best understood by the worker? Or which is the more likely to be read with attention?

Consider the contents of the two pages. The first contains 278 words. The revised page contains 100 words. A glance will show that the difficulty level of the words used in the revised page is far below that of the first sample. It is far easier to read the revised

page with its abundance of white space. And the drawing and "Reward" box, simple as they are, are attention catching features.

It may be argued that the revised page has omitted some of the information contained in the original. That is true. The essential information is there, however. It must be borne in mind when comparing these two pages that a large number of the employees who were to receive this handbook had only an eighth grade education. The writer believes that it is better to present the essential idea in a way that will get it across than to present a large amount of information which may not even be read.

A comparative analysis of the original and revised handbooks was made, using Flesch's new Readability Formulae (4). Thirteen 100-

word samples were drawn from the original, and fourteen 100-word samples were drawn from the revised handbook. The results of the application of the formulae to these samples are shown in Tables 1 and 2.

The average Readability Score for the original handbook was 59.2, a difficulty level requiring some high school for adequate understanding. The scores ranged from "Difficult" to "Fairly Easy" with 8 of the 13 scores "Fairly Difficult" or harder. This means that over half of the samples drawn from the original handbook required "some high school" or more education for adequate understanding. In view of the fact that a large number of the employees who were to receive the handbook had eighth grade or less education, it is evident that its effectiveness would have been limited.

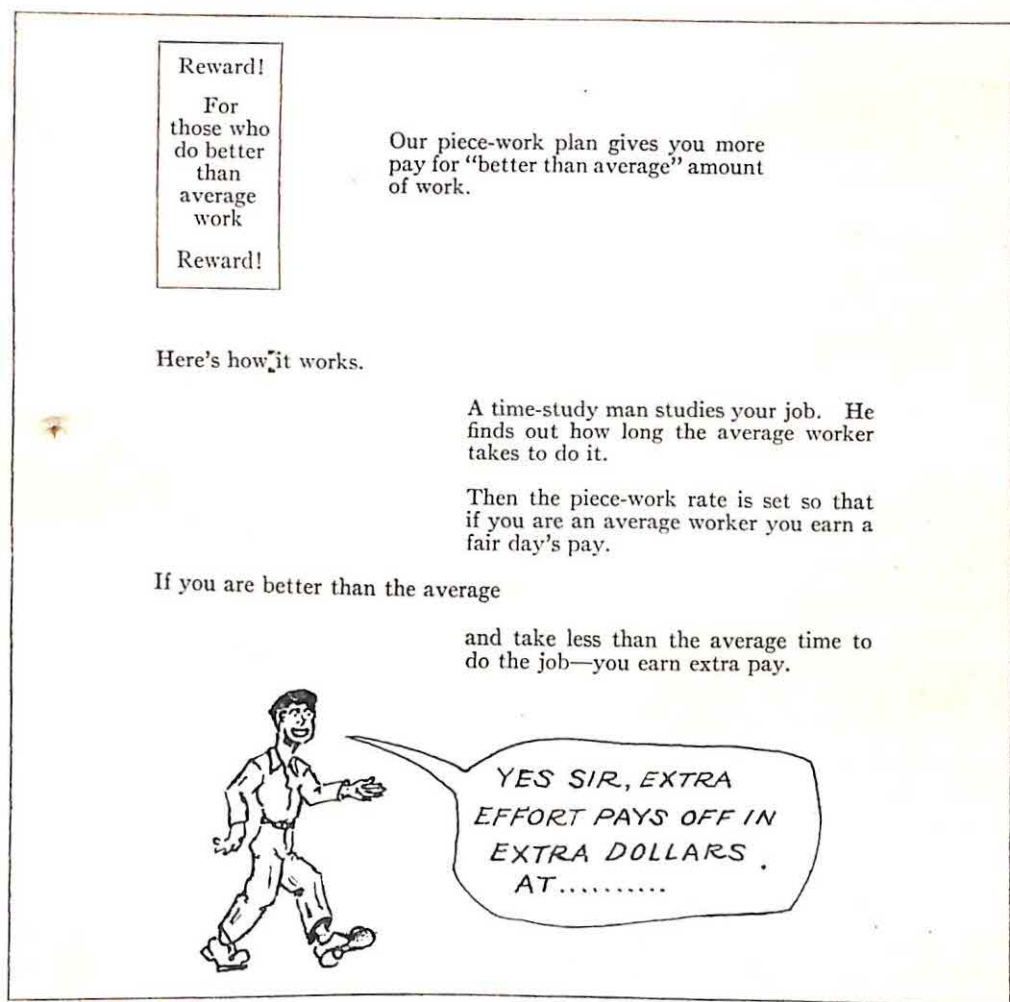


Fig. 2. Sample page from revised handbook.

Table 1

Flesch Readability Scores and Interpretation for Samples Drawn From Original and Revised Handbooks

Readability Score	Tabulation		Style	Typical Magazine	School Grade Level of Potential Audience
	Original	Revised			
0 to 30			Very difficult	Scientific	College
30 to 50	3		Difficult	Academic	H.S. or some college
50 to 60	5		Fairly difficult	Quality	Some H.S.
60 to 70	2		Standard	Digests	7th or 8th grade
70 to 80	3	2	Fairly easy	Slicks—fiction	6th grade
80 to 90		7	Easy	Pulp—fiction	5th grade
90 to 100		5	Very easy	Comics	4th grade
Average	59.2	87.7			

Table 2

Flesch Human Interest Scores and Interpretation for Samples Drawn From Original and Revised Handbook

Human Interest Score	Tabulation		Description of Style	Typical Magazine
	Original	Revised		
0 to 10			Dull	Scientific
10 to 20	2		Mildly interesting	Trade
20 to 40	3	2	Interesting	Digests
40 to 60	5	4	Highly interesting	New Yorker
60 to 100	3	8	Dramatic	Fiction
Average	45.9	59.3		

The average Readability Score for the revised handbook was 87.7, a difficulty level requiring only a fifth grade education for adequate understanding. The scores for the samples ranged from "Fairly Easy" to "Very Easy." Very few employees would be unable to understand material written at this level.

The Human Interest scores were also higher for the revised handbook, as shown in Table 2. This means that the revised handbook contained more personal pronouns and personal sentences than did the original. This, according to Flesch, makes the material more interesting to read.

Summary

1. Recognition of the importance of communication between management and workers has resulted in an increase in the use of

written media such as house organs, handbooks, etc. Discussions of the content, purposes, and uses of these media are available in the literature. Discussions of the technical problems of publication are also available.

2. The important problem of readability in these written communications has not been dealt with, however. There is little in the communications literature dealing with the need to write at the level the worker can understand, and little dealing with the problem of how to write readable material.

3. The Readability Formulae of Flesch now provide an objective method of evaluating the difficulty level of written material, and the rules laid down by Flesch point the way to writing readable copy.

4. Application of the Readability Formulae to management-employee communications indicates that they are frequently written at a

difficulty level far too high for the intended audience.

5. Application of the formulae to a proposed employee handbook classified it as "Fairly Difficult," requiring "some high school" reading ability for adequate comprehension. Many of the workers who were to read the handbook, however, had only eighth grade or less education. The handbook was rewritten, using short sentences, short common words, and more personal words. Excess words were discarded, leaving more white space per page, thus making it easier to read. Flesch analysis classified the material as "Easy," requiring fifth grade reading ability for understanding.

Received April 16, 1949.

References

1. Bentley, G. *How to edit an employee publication*. New York: Harper & Brothers Publishers, 1944.
2. Biklen, P. F., and Breth, D. B. *Successful employee publication*. New York: McGraw-Hill Book Company, Inc., 1945.
3. Flesch, R. *The art of plain talk*. New York: Harper and Brothers, 1946.
4. Flesch, R. A new readability yardstick. *J. appl. Psychol.*, 1948, 32, 221-233.
- 4a. Flesch, R. *The art of readable writing*. New York: Harper and Brothers, 1949.
5. Heron, A. R. *Sharing information with employees*. Stanford: Stanford University Press, 1942.
6. American Management Association. *How to prepare and publish an employee manual*. New York, 1946.
7. Paterson, D. G., and Jenkins, J. J. Communication between management and workers. *J. appl. Psychol.*, 1948, 32, 71-80.



Reliability of the Flesch Readability Formulas

Patricia M. Hayes, James J. Jenkins, and Bradley J. Walker *

Department of Psychology, University of Minnesota

A formula developed by Flesch (3) for estimating the comprehension difficulty of written material has received widespread attention in many areas of communication. It has been applied in the fields of journalism (7), advertising (1), industrial communications (5), government publications (8) and many others.

In view of the wide use of his formula, Flesch (4) published a revision in 1948 designed to increase its utility and make its application and interpretation easier. This revision proposes the use of two formulas to measure two relatively independent aspects of readability. The first formula involves word length (*wl*) and sentence length (*sl*) and gives a measure of "reading ease" (RE). The second formula, based on personal words (*pw*) and personal sentences (*ps*), yields a measure of "human interest" (HI).

As these formulas become increasingly popular, they must, of course, be evaluated critically. Like other psychological tools, they must be tested for validity and reliability. Flesch (4) reports several studies of the validity of the original formula which indicate that material rated more readable by the formula also proves more readable in terms of readership surveys and opinions of judges. As yet, however, no studies of the reliability of the formulas as applied by different analysts have been reported.

The following studies were designed as first steps in the examination of analyst-to-analyst reliability of the formulas to determine the extent to which they are effectively objective.

First Study

The material chosen for analysis in the first study consisted of the 40 prize-winning letters in the recent General Motors' "Why I Like My Job" contest (9). These letters were selected because they presented a wide range

* Hayes and Jenkins are primarily responsible for the first study in this paper and Walker is primarily responsible for the second.

of difficulty, style, structure and content. It was believed the letters would afford a maximum number of problems in interpretation and would provide a rigid test of objectivity.

Two sets of samples were drawn from the 40 letters. Each set consisted of two 100-word samples from each letter. Since the letters ranged in length from about 350 to 3000 words with a median length of 750 words, there was little overlapping between the sets of samples.

Two experienced and two inexperienced analysts participated in the study.¹ The experienced workers analyzed both sets of samples; the inexperienced each analyzed one set. The experienced analysts had worked with the original formula and the revised formulas for a year and a half. The inexperienced analysts had never worked with the formulas before. The analysts made no attempt to agree on interpretation of Flesch's instructions and refrained from discussing interpretation with anyone else.

Reading ease and human interest scores were computed from tables (2) in an effort to minimize computational errors. Results of analyses made by different investigators on the same set of samples were compared by determining: the significance of differences between means of the four variables (*wl*, *sl*, *pw*, *ps*) and the two scores (HI, RE) for the set of samples as a whole; the extent of correlation between results of analysts; the number and degree of differences in actual scores for each sample; and the number of differences in descriptive categories assigned to each sample.

Results of First Study

The assumption of wide variability in the material used was confirmed. As may be seen in Table 1, samples ranged from 6 to 88 on the reading ease scale of 0 to 100 points and

¹ The writers would like to acknowledge the assistance of Barbara Lee and James Farr in this part of the project.

Table 1

Means, Standard Deviations and Ranges for the Four Variables and Two Scores of the Flesch Readability Formulas for Each Analyst

Mean	Analyst	<i>wl</i>	<i>sl</i>	RE	<i>pw</i>	<i>ps</i>	HI
Set 1	A	145.0	23.7	60.2	9.9	15.3	41.1
	B	145.2	23.3	60.4	9.7	10.2	38.2
	C	145.0	23.9	59.9	9.5	13.6	39.5
Set 2	B	144.0	22.7	62.1	9.5	11.8	38.6
	C	143.8	22.5	62.3	10.2	14.2	42.2
	D	144.3	23.2	61.2	9.5	13.1	39.3
Standard deviation							
Set 1	A	11.9	11.1	15.9	3.5	26.6	18.2
	B	11.8	11.3	16.1	3.3	17.6	16.4
	C	11.7	11.1	15.4	3.7	18.4	15.6
Set 2	B	9.4	10.8	13.6	2.9	21.4	14.3
	C	9.0	9.8	12.9	2.8	26.7	13.6
	D	9.7	9.9	13.4	3.4	21.1	14.0
Range							
Set 1	A	124-174	11-72	7-85	4-20	0-100	15-100
	B	125-174	10-72	6-84	4-20	0-83	13-97
	C	124-173	11-72	6-82	5-18	0-91	18-94
Set 2	B	122-164	12-80	42-81	5-15	0-100	16-79
	C	122-162	11-74	42-88	5-17	0-100	18-80
	D	123-164	12-74	42-86	5-15	0-100	16-80

from 13 to 100 on the human interest scale of 0 to 100 points.

The means of the four variables and two scores obtained by analysts on the same sets of samples were tested for significant differences by use of the critical ratio corrected for correlation. None of the differences between any analysts in either sample set proved to be significant at the five per cent level.

Rank difference correlations were computed between each pair of analysts within each sample set on the rank given each letter. These correlation coefficients are presented in Table 2. All of the correlations are positive and significantly different from zero beyond the one per cent level.

An inspection of Table 2 indicates that analysts were in good agreement in interpreting the components and final score for reading ease. For the human interest variables, however, there was much less agreement between analysts. Personal words were apparently in-

terpreted much the same by analysts, but it is evident there were diverse interpretations of personal sentences. This acts, of course, to lower the correlations of the human interest score.

It should be noted that correlations between experienced analysts (B and C) are not appreciably different from those with inexperienced analysts (A and D).

Since neither of the statistical methods presented above reveals the actual differences between analysts for a given sample, a third kind of comparison was made. Within each sample set all analysts were compared on reading ease and human interest scores for each letter. Actual point differences for each pair of analysts were tabulated. Results of the 240 comparisons are shown in Table 3.

This table also suggests that there is greater agreement on reading ease (90 per cent of the comparisons within four points of each other) than on human interest (90 per cent of the comparisons within eight points of each other). On the 100-point scale designed to be used as an estimating device, deviations as small as these do not appear to be of great importance.

Again it should be mentioned that no consistent difference was found in the number or extent of deviations between scores of experi-

Table 2

Rank Order Correlation Coefficients between Pairs of Analysts for the Variables and Scores of the Flesch Readability Formulas*

Analysts**	Sample Set 1			Sample Set 2		
	A and B	A and C	B and C	C and D	B and D	B and C
<i>wl</i>	.99	.99	.99	.99	.99	.99
<i>sl</i>	.94	.97	.94	.83	.87	.94
RE	.98	.99	.98	.93	.95	.97
<i>pw</i>	.93	.94	.96	.92	.99	.93
<i>ps</i>	.89	.69	.74	.63	.87	.60
HI	.88	.91	.96	.78	.97	.80

* These correlations should be interpreted with caution since the data are markedly skewed in the case of *ps* and HI. It should be noted, however, that the order of relative accuracy for the scores is the same whether correlations, point differences or category differences are considered.

** Analysts A and D are inexperienced; B and C, experienced.

enced analysts compared to their deviations with scores of inexperienced analysts.

A final comparison was made between analysts in terms of descriptive categories in which results are often reported and utilized. Flesch (4) divides the reading ease range into seven levels varying from "very difficult" to "very easy" and the human interest range into five levels varying from "dull" to "dramatic."

Table 3

Differences in Score Points Between Analysts on Identical Samples *

Difference in Points	Reading Ease		Human Interest	
	Per Cent of Comparisons	Cumulative Percentage	Per Cent of Comparisons	Cumulative Percentage
1	49.2	49.2	33.3	33.3
2	25.8	75.0	18.0	51.3
3	9.2	84.2	6.6	57.9
4	6.2	90.4	15.4	73.3
5	3.8	94.2	4.2	77.5
6	1.6	95.8	5.4	82.9
7	1.7	97.5	3.4	86.3
8	.8	98.3	3.7	90.0
9	.9	99.2	2.5	92.5
10	.8	100.0	.8	93.3
11 or more			6.7	100.0

* Based on 240 comparisons; three analyses for each of 80 samples of 200 words.

Of the 240 comparisons, in only 14 cases (5.8 per cent) did the analysts differ in the category assigned to reading ease. In 28 cases (11.7 per cent) they differed in the category assigned to human interest. Of these cases of disagreement, only two (.8 per cent) were greater than one category for reading ease, and only four (1.7 per cent) were greater than one category for human interest.

Discussion

While the results of this first study seem to constitute a limited but fairly clear answer to the question of reliability of the Flesch formulas, mention of the greatest sources of error may be of some value in interpreting the data and may provide a few hints to those who wish to use the formulas.

The greatest discrepancies obviously appear in interpretation of personal sentences. A study of Table 2 shows that correlations for this variable are especially low when analyst "C" is involved. Samples which contributed most to the discrepancy between "C" and the other analysts were studied. Over half of the major differences between "C" and the others involved one type of personal sentence defined by Flesch (4) as "grammatically incomplete sentences whose full meaning has to be inferred from the context." The examples given by Flesch (4) appear to be taken from conversations, and apparently the definition was regarded as limited to conversations by analysts "A," "B" and "D." It would seem, however, that the examples are to this extent misleading since conversational sentences are already covered by Flesch's first definition regarding spoken sentences. It might be suggested that analysts study the definitions carefully and that Flesch provide more varied examples.

A second source of disagreement involved rhetorical questions. Analyst "C" did not count these as personal sentences, but it appears clear from Flesch's definition that these should have been considered and scored as the other analysts scored them.

If these two sources of error (incomplete sentences and rhetorical questions) had been corrected, correlations for personal sentences would have been raised above .90, and the human interest scores would have appeared much more reliable.

Errors in personal words were few and appear to be due largely to carelessness. While there were no consistent errors, from time to time one analyst or another tended to regard a common-gender noun like "worker" or "manager" as a personal word. Flesch's definitions and examples are explicit on this point (4).

Errors in sentence length resulted chiefly from disregarding directions on counting the last sentence when the sample ends in the middle of a sentence and from disagreement on breaking sentences into units of thought. It may be noted that the lowest correlations in Table 2 for sentence length involve analyst "D." A study of the samples yielding the greatest discrepancies revealed that if analyst "D" had broken sentences into units of thought in just

two instances, correlations would have been above .95. Here it appears that more careful attention to directions would have assured high reliability.

Errors in word length are all very small and appear to reflect minor clerical errors.

Second Study

A second study was conducted to test our findings with a large number of inexperienced analysts. Samples of 500 words from 63 house organs and employee publications which were being examined in connection with a continuing study of industrial communication (6) were assigned for analysis to 18 members of a graduate seminar in psychology. Each student analyzed seven publications which were subsequently reanalyzed by another member of the seminar. Assignments were anonymous and cooperation between students was discouraged. Only three of the students had appreciable experience with the formulas prior to the time of the study.

Table 4

Sampling Statistics of Test and Re-Test Distributions for the Second Study

	Means		Standard Deviation		Range	
	Test	Re-Test	Test	Re-Test	Test	Re-Test
<i>wl</i>	155.4	154.9	7.98	7.98	138-172	140-167
<i>sl</i>	20.6	20.5	4.39	4.15	15-45	13-45
<i>RE</i>	54.5	55.2	8.40	7.71	30-73	31-69
<i>pw</i>	7.3	6.6	2.38	2.24	3-16	2-13
<i>ps</i>	12.8	13.1	11.17	11.45	0-46	0-48
<i>HI</i>	30.3	28.3	10.74	9.56	15-64	8-62

The results of the analyses provided pairs of scores for each of the publications. The first analyses were compared with the second analyses to determine the reliability of the application of the formulas to the same samples. Product moment correlations between the "test" and "retest" analyses are as follows: *wl*, .90; *sl*, .92; *RE*, .91; *pw*, .78; *ps*, .64; and *HI*, .81. All coefficients were positive and significantly different from zero.

The data for the means, standard deviations and ranges are presented in Table 4. A comparison of the standard deviations and the

ranges in Table 4 with those in Table 1 reveals that the material used in the second study was appreciably more homogeneous than that used in the first. The correlations found in the second study, then, might be expected to be smaller than those of the first study. Accordingly, the correlations presented immediately above were "corrected" by estimating their magnitude on the basis of the more heterogeneous material of the first study. This "correction" gave the following coefficients: *wl*, .95; *sl*, .99; *RE*, .98; *pw*, .88; *ps*, .85; and *HI*, .92.

These correlations approximate those found in the first study and would lead one to the same conclusions. Reading ease with its components is analyzed quite reliably and human interest with its components is analyzed with less, though still fair, reliability. Analysis of personal sentences again shows the greatest lack of agreement between analysts.

The difference in points between "test" and "retest" analyses agrees rather closely with the data from the first study given in Table 3. Ninety per cent of the paired scores for reading ease were within six points of each other and approximately ninety per cent of the paired scores for human interest were within eight points of each other.

Summary and Conclusions

An examination of analyst-to-analyst reliability of the Flesch readability formulas was conducted. In the first study two sets of samples were drawn from reading material of a highly variable nature believed to involve a large number of problems of interpretation. The sets of samples were analyzed by two inexperienced and two experienced analysts. Results of analysts for each set of samples were compared by testing the significance of mean differences on the variables and the scores, correlating results on the variables and scores, tabulating deviations in terms of score points and tabulating disagreements in descriptive categories.

In the second study, eighteen students analyzed samples of 500 words from 63 industrial house organs. Each sample was independently analyzed by two analysts. Correlations between the first and second sets of

analyses were computed and then corrected for restriction of range. Deviations in terms of score points were computed.

From the above data the following conclusions seem justified:

1. Analyst-to-analyst reliability on word length, sentence length, and reading ease is quite high for the kinds of material used in this study.

2. Analyst-to-analyst reliability on personal words is fair, but on personal sentences (and as a result on human interest) is lower than might ordinarily be considered desirable.

3. For practical purposes the Flesch formulas and the directions for their use are sufficiently objective to be used even by inexperienced analysts to obtain estimates of the reading ease and human interest of written material.

Received June 8, 1949.

References

1. Alden, J. Lots of names—short sentences—simple words. *Printer's Ink*, June 29, 1945, 21-22.
2. Farr, James N., and Jenkins, James J. Tables for use with the Flesch readability formulas. *J. appl. Psychol.*, 1949, 33, 275-278.
3. Flesch, R. *The art of plain talk*. New York: Harper and Brothers, 1946.
4. Flesch, R. A new readability yardstick. *J. appl. Psychol.*, 1948, 32, 221-233.
5. Paterson, D. G., and Jenkins, James J. Communication between management and workers. *J. appl. Psychol.*, 1948, 32, 71-80.
6. Paterson, D. G., and Walker, B. J. Readability and human interest of house organs. *Personnel*, 1949, 25, 438-441.
7. Swanson, C. E. Readability and readership: a controlled experiment. *Journ. Quart.*, 1948, 25, 339-343.
8. *How does your writing read?* U. S. Civil Service Commission. Washington: U. S. Government Printing Office, 1946.
9. *The worker speaks*. General Motors, 1947.

The MacQuarrie Test for Mechanical Ability.

IV. Time and Motion Analysis *

Charles H. Goodman

Radio Corporation of America

This is the fourth¹ and last article on the use of the MacQuarrie Mechanical Ability Test in a radio manufacturing company. This study presents the results of a time and motion analysis of the test. The purpose of this study was to determine whether a time and motion analysis could provide insight into the number of factors being measured by the test, and whether this type of analysis could assist in the interpretation of the factors in the test.

Of the seven MacQuarrie sub-tests, only four—Tracing, Tapping, Dotted, and Copying—lend themselves to time and motion analysis. The reason is that only these four sub-tests involve manual movements and could therefore be analyzed by this technique. The particular method of time and motion analysis used in this study was that of *synthesis*,² which is defined as a method of determining the select time for a given motion pattern by the application of standard moving time values to a detailed motion analysis.

Procedure

Upon completion of the time and motion analysis, it was possible to construct Table 1 which shows the various time and motion elements found to be operating in the four sub-tests. This analysis also provided the data whereby one could determine which elements were common among the tests and the frequency with which these elements occurred while carrying out the tasks.

* The writer wishes to thank Mr. Fred Weber, Time Study Engineer, RCA Victor Corporation, for his assistance in conducting the time and motion analysis for this study.

¹ Goodman, Charles H. The MacQuarrie test for mechanical ability. I. Selecting radio assembly operators. *J. appl. Psychol.*, 1946, 30, 586-595; II. Factor analysis. *J. appl. Psychol.*, 1947, 2, 150-154; III. Follow-up study. *J. appl. Psychol.*, 1947, 5, 502-510.

² *Synthesis manual*. Radio Corporation of America, RCA Victor Division, Camden, N. J., 1944.

Results

Overlapping *r*'s were computed on the basis of the data recorded in Table 1. The formula used to compute the overlapping *r*'s was:

$$r = \frac{nc}{\sqrt{(na+nc)(nb+nc)}}$$

The object of computing these overlapping *r*'s was to determine whether the relationships among these four tests as based upon the time-study elements, would approximate the relationship among the same tests as shown by the Pearsonian intercorrelations. Table 2 shows the overlapping *r*'s and their Pearsonian counterparts. Four of the six overlapping *r*'s deviate from their Pearsonian *r*'s from .00 to .09. The overlapping *r* for the Tracing and Tapping Test deviates .21 from its corresponding Pearson *r*. The sixth and largest deviation of .23 involves the Tapping Test with Copying. The writer is unable on the basis of available data to explain why these two deviations are larger than the other four.

While it appears that there is considerable agreement between the Pearsonian *r*'s and the overlapping *r*'s, one cannot lose sight of the fact that in using the formula for computing the overlapping *r*'s one does not know if the factors operate independently and additively, or in a dependent and related manner. While this finding is of considerable interest, further study would certainly be needed before one could attach any significance to it.

Since it was found that considerable agreement existed between the Pearsonian *r*'s and the overlapping *r*'s, the question was raised as to whether or not time and motion analysis would be helpful in interpreting the factors being measured by the MacQuarrie test.

An earlier study³ had been made by the

³ Goodman, C. H. *Op. cit.*

Table 1

Time Study Elements Operating in the Tasks Required by the MacQuarrie Tests

	Transport Loaded	Inspection	Simo Inspection	Control	Definite Stop	Align	Finger Wrist Movement	Finger Wrist Forearm Movement	Finger Wrist Forearm Upper Arm Movement	Elbow Pivot	Shoulder Pivot
Tracing	x	x		5x	x	2x	x				
Tapping	x		x	x				x		x	
Dotting	2x		x	4x	2x	x			x	x	x
Copying	x	8x		2x	x			x	x	x	

writer in an endeavor to identify the factors in the MacQuarrie test by means of factor analysis. This factorial study showed three factors to be operating in the MacQuarrie test. The test loadings of the first factor ranged from .369 to .000. While none of the test loadings for this factor were statistically significant,⁴ it appeared to the writer that some factor was operating. The sub-tests of Tracing, Dotting, and Pursuit had loadings of .369, .338 and .325 respectively, while the remaining four sub-tests were zero or slightly above zero. While the writer was unable to identify the factor on the basis of the factor analysis data, time and motion analysis data clearly indicated to the writer that it was a "visual inspection" factor, since the time and motion analysis showed that the three largest test loadings on this first factor called for considerable amounts of visual inspection before the tasks could be done. On the other hand, little, if any, visual inspection was called for by the remaining four sub-tests. From a statistical point of view this factor was named only tentatively since none of the loadings met the criteria of being as large as .40. On the other hand, based on the time study analysis, the writer had no hesitancy in so naming the factor. It is further the opinion of the writer that he could not have readily identified this factor without the time study data.

The test loadings of the second factor in the factorial analysis were statistically significant

and this factor was named a "spatial" factor. Corroborating evidence that this factor was a "spatial" factor comes from the work of Thurstone⁵ and Harrell,⁶ in that their studies also showed the MacQuarrie to be loaded with a "spatial" factor.

The time and motion analysis, in the opinion of the writer, did not in this instance provide any assistance in identifying this "space" factor. The reason for this may be that the space factor is psychological in content, presumably involves mental processes and therefore escapes the detection of time and motion studies, a technique which is only applicable to observable acts.

In naming the third factor disclosed by the factorial analysis, the writer found the time and motion data of material assistance. It was apparent from such analysis that the tests carrying the highest loadings involved manual movement. This might readily have been concluded from mere study of the tests themselves. However, further study of the data produced by the time study analysis showed that it was more than mere manual movement. The time study data showed that in carrying out these tasks it was necessary to exert careful control in carrying out the manual movements. This control involved the muscles being used in order that motions could be stopped when there was need to make alignment and to carefully guide the pencil in prescribed areas.

⁴ Thurstone, L. L. *Primary mental abilities*. Chicago: University of Chicago Press, 1938.

⁵ Thurstone, L. L. *Op. cit.*

⁶ Harrell, W. A factor analysis of mechanical ability tests. *Psychometrika*, 1940, 5.

Further evidence to support this view was obtained from high speed camera photographs which showed the motion in operation and the control being exercised. It was on the basis of this evidence that the writer named this factor a "controlled manual movement."

Harrell,⁷ in his studies with the MacQuarrie, found a factor evidently similar to that of the writer. Harrell named this factor a "manual agility" factor. On the basis of the foregoing evidence, the present writer cannot accept the concept and connotation of this factor as being a "manual agility" factor.

Table 2

Pearsonian Correlations Compared with Correlations
Computed on the Basis of Overlapping
Elements

	Tracing		Tapping		Dotting	
	Pear- son <i>r</i>	Over- lapping <i>r</i>	Pear- son <i>r</i>	Over- lapping <i>r</i>	Pear- son <i>r</i>	Over- lapping <i>r</i>
Tracing						
Tapping	.48	.27				
Dotting	.55	.58	.47	.50		
Copying	.44	.39	.31	.54	.34	.43

As a result of this study it would appear to the writer that the technique of time and motion analysis might be of considerable assistance in helping to analyze performance tests for the purpose of identifying the factors involved in such tests. It would seem that the area of mechanical dexterity would be a fruitful field to explore. One should not lose sight of the fact that time and motion analysis will probably tend to show a greater number of elements in operation than does the statistical factorial method. This can be seen in Table 2 of the present study where some eleven elements are shown as compared with the three factors disclosed by factor analysis. This may be due primarily to the fact that the elements

of time and motion analysis are more minutely broken down and that there may be various combinations of these elements in a single task. However, there is no available evidence to prove or disprove this contention. More important is the fact that time study elements should not be considered as being equal or comparable to the so-called psychological factors disclosed by factor analysis.

Summary

This study has presented the findings of a time and motion analysis of four of the subtests of the MacQuarrie Mechanical Ability Test. The purpose of the study was to determine whether time and motion analysis could assist in the interpretation of the factors being measured by the test.

On the basis of the findings of this study the following conclusions might be warranted:

1. Correlations computed on the basis of overlapping time and motion elements give, in four of the six cases, *r*'s closely similar to their Pearsonian *r*'s. Deviations in four cases range from .00 to .09. The fifth and sixth deviations were .21 and .23. It would be mere speculation on the part of the writer to explain these larger deviations.

2. There is close agreement in the analysis of the overlapping elements as produced by the factor analysis and by the time and motion analysis. The factor analysis identified a space factor, and a manual movement factor. The time and motion analysis revealed a visual inspection factor and a controlled manual movement factor. The space factor, which appears to be more psychological in nature, was not revealed by the time and motion analysis.

3. Time and motion analysis technique appears to have the possibility of being a valuable adjunct for gaining analytical insight into psychological tasks involving manual movements.

⁷ Harrell, W. *Op. cit.*

The Pre-Engineering Inventory as a Predictor of Success in Engineering Colleges *

Frederic Lord and John T. Cowles

Educational Testing Service

and

Manuel Cynamon

Brooklyn College

The Pre-Engineering Inventory is a battery of seven objective tests, designed primarily to assist in the selection of those students who will be most likely to succeed in engineering schools.¹ This report summarizes a group of related studies on the reliability of this battery of tests and on their predictive efficiency, as evidenced particularly by correlations of the test scores with various measures of engineering school achievement.

The Pre-Engineering Inventory is intended for use both as an instrument for selecting entering students and for the guidance of students before or during the first two years of undergraduate engineering studies. It is the purpose of the *Pre-Engineering Inventory* to supply measures of only those aptitudes or background factors which lend themselves to reliable measurement by means of written tests. Admissions and guidance officers are advised to use *Pre-Engineering Inventory* test scores in conjunction with other information about the student obtained from previous academic rec-

ords, recommendations by teachers, and personal interviews.

Description of Tests and Scores

The present form of the *Pre-Engineering Inventory*, known as Revised Form A, was first made available in 1944, and has been used in two principal types of testing programs: a national program administered on fixed dates at widespread testing centers, and an institutional program administered on varying dates by cooperating institutions. The national program was designed for testing applicants for entrance to engineering schools, the institutional program for testing students already enrolled in those institutions.

The difficulty level of the tests of the *Pre-Engineering Inventory* has been adjusted to the range of capacity of a base sample of pre-war freshmen in a broadly representative group of accredited engineering schools; the present form of the Inventory is not intended to be used in high schools or liberal arts colleges.

The tests of the *Pre-Engineering Inventory* are contained in two booklets, with separate answer sheets adapted to machine or hand scoring. The questions are of the objective, multiple-choice type and include passages requiring interpretation, problems for solution, diagrams to comprehend, and questions on specific information. Emphasis is placed on essential aptitudes, skills, fundamental information, and the discovery or application of principles, rather than on mere factual memory or achievement in specific courses of instruction. The seven tests included in the battery

* The authors have been aided in the preparation of this article by the valuable suggestions and criticisms of Dr. A. Pemberton Johnson, Dr. William G. Mollenkopf, Dr. William B. Schrader, and Dr. Ledyard R. Tucker. We especially wish to acknowledge the able editorial work of Mrs. Mabel K. Rugg.

¹ The Inventory is one of several related groups of tests now in use or in process of development as part of the Measurement and Guidance Project in Engineering Education. This Project, initiated in 1943, is sponsored by the American Society for Engineering Education and the Engineers' Council for Professional Development. The testing and research activities of the Project were transferred from the Graduate Record Office to the Educational Testing Service in January 1948 at the time the latter organization was established. An advisory council representing the two sponsoring societies meets periodically with representatives of the Educational Testing Service in order to review and guide the policies and progress of this Project.

Table 1

Reliabilities of Raw Scores on the Pre-Engineering Inventory Tests for Five Schools Participating in the Fall 1947 Testing and for a Random Sample from the April 1948 National Program

Tests	School					National Program
	V	W	X	Y	Z	
I. General Verbal Ability (100 items)	.93	.94	.93	.93	.94	.94
II. Technical Verbal Ability (90 items)	.92	.94	.90	.85	.91	.92
III. Ability to Comprehend Scientific Materials (100 items)	.91	.88	.89	.89	.91	.88
IV. General Mathematical Ability (90 items)	.90	.91	.89	.85	.85	.89
V. Ability to Comprehend Mechanical Principles (52 items)	.87	.91	.75	.84	.75	.77
VI. Spatial Visualizing Ability (56 items)	.89	.88	.91	.90	.90	.92
VII. Understanding of Modern Society (70 items)	.81	.83	.76	.82	.86	.80
Number of Cases	178	128	172	181	98	366

are:² Test I, General Verbal Ability; Test II, Technical Verbal Ability; Test III, Ability to Comprehend Scientific Materials; Test IV, General Mathematical Ability; Test V, Ability to Comprehend Mechanical Principles; Test VI, Spatial Visualizing Ability; and Test VII, Understanding of Modern Society.

In addition to the separate scores for the seven tests of the Inventory, an eighth score, the *Composite Score*, is obtained by adding together the raw scores of Tests II, III, and IV. This *Composite Score* represents a general verbal and quantitative aptitude score. Preliminary studies on experimental forms of the test indicated that, for practical purposes, this *Composite Score* was the best index obtainable from the entire battery of the candidate's general aptitude for engineering study; it was one of the purposes of the present study to reexamine this preliminary finding concerning the *Composite Score*.

² See K. W. Vaughn, *The Pre-Engineering Inventory*, *J. Engng. Educ.*, 1944, 34, 615-625, for a description of the separate tests and the criteria by which they were chosen.

Reliability of the Tests

Reliability coefficients are given in Table 1 for a random sample group taken from the April 1948 nationwide testing of applicants to engineering schools, and also for recently enrolled freshmen students in five schools where testing was carried out during the fall of 1947. These coefficients are based on a division of each test into "rational halves." In making this division any group of items based on a single paragraph of reading matter or on a single chart or table was treated as an indivisible unit. An attempt was made to match the halves, in so far as possible, with respect to item difficulty, subject matter, item type, and position in the test.

The reliability coefficients were obtained from the correlation between half-test scores by the application of the Spearman-Brown Formula (modified so as to take into account the fact that the two "halves" of the test did not necessarily contain equal numbers of items). Since the number of groups of items in a single test was not always large, it was often impossible to achieve satisfactory match-

ing of the test items. As a result, the reliability coefficients presented probably tend to be underestimates of the actual test reliabilities in many cases.

Test V (*Ability to Comprehend Mechanical Principles*) and Test VII (*Understanding of Modern Society*) do not have as consistently high reliabilities as would be desired. The level of reliability of the other five tests appears to be sufficiently high to justify a good measure of confidence in the scores. It may be stated that there is a strong tendency in these data for a test to have the highest reliability coefficients in those groups having the highest mean or the greatest variability of scores. (It would be desirable both in Table 1 and in most of the subsequent tables to present the mean and standard deviation of each of the different groups studied on each of the eight Inventory scores, but space considerations make it undesirable to present all relevant data. The schools and other groups studied here and throughout this report differ very greatly among each other with respect to mean score and with respect to variability of scores on any one test.)

Since the *Composite Score* is used much more than the score on any single test, its reliability is of greater importance than those presented in the table. The *Composite Score* reliability for the 366 cases in the April 1948 National Program group is 0.95. This is a highly satisfactory value. The comparable coefficients for the different school groups in Table 1 have not been calculated, but the general level of the test reliabilities indicates that the *Composite Score* reliabilities for the several schools could be expected to fluctuate in the neighborhood of 0.95.

Twelve-School Study³

Studies were made of the correlations of Inventory scores with achievement records and other relevant data for the freshman and, in some cases, the sophomore classes in twelve

³ This twelve-school study was planned by K. W. Vaughn as Director of the Graduate Record Office and was carried through largely under his direction. For earlier validity figures see K. W. Vaughn. Basic considerations in a program of freshman evaluation. *J. Engng. Educ.*, 1944, 35, 161-179.

engineering colleges.⁴ The test scores for these students had been obtained by testing the entire freshman class shortly after enrollment. The names of the institutions are as follows: California Institute of Technology; Carnegie Institute of Technology; Columbia University; Georgia School of Technology; Massachusetts Institute of Technology; Newark College of Engineering; North Carolina State College of Agriculture and Engineering; Oklahoma Agricultural and Mechanical College; Oregon State College; University of California at Los Angeles; University of Michigan; and University of Texas. In two cases, two different groups of students from the same school were analyzed separately because they took dissimilar courses, so that fourteen, rather than twelve, "school groups" are discussed in the present report. In reporting results the identity of the schools will not be disclosed. Dates of testing—which cover the period from 1944 to 1946—and also numbers of individuals will be presented with the other data in the tables which follow.

Care should be taken in considering the correlation or lack of correlation of the *Pre-Engineering Inventory* scores with school grades. A score on one of the Inventory tests may have little usefulness for predicting average school grades but may nevertheless prove to be very useful for guidance or for selecting individuals with good potentialities for success after graduation from engineering school. The *Understanding of Modern Society* test, for example, should not be expected to correlate highly with engineering school average grades.

Another fact that must be taken into account is that school grades are rarely as statistically reliable as the scores on objective tests. Hence correlations with school grades often present attenuated estimates of relationship. Moreover the necessity of utilizing only those students who actually were admitted to engineering school and persisted through one or more semesters of study limits the range of talent to an upper range, as compared with the total range of talent in the group from which

⁴ A year after the study reported here was begun, data on the academic achievement of students tested as freshmen were obtained from twenty-two additional engineering colleges. These data have not been analyzed to date.

selection was originally made. This definitely reduces the amount of correlation which is obtained, particularly if the degree of selection is great on those characteristics measured by the selection test. The present study, therefore, is limited to demonstrating how well the *Pre-Engineering Inventory* correlates with the engineering school grades of students already admitted to engineering school. It does not demonstrate how well the *Pre-Engineering Inventory* selects from among all engineering applicants those who will obtain the best engineering school grades, nor how well the *Inventory* would predict an ideal measure of engineering school success above and beyond course grades, including the examinee's later work as a practicing engineer.

1. *Correlations of the Test Scores with Average Grades.* One of the main purposes of this study was to determine the validity of the *Composite Score* for predicting engineering school success as measured by the various average grades available for enrolled students. The

Composite Score validities were obtained by computing the correlations between the *Composite Score* and the various available averages of course grades. The correlations between scores on the individual *Pre-Engineering Inventory* tests and average grades were also obtained in order to determine how scores on the separate tests might be related to engineering school success. These latter correlations should not be considered as validity coefficients since it is not intended that any single test score should be used to predict average grades in engineering school.

The correlations of each test score and of the *Composite Score* with average first-term grades are presented in Table 2; the *Composite Score* validities for school groups having more than one term of achievement records are presented in Table 3. It will be noted that Table 2 shows a median correlation of 0.60 between *Composite Score* and average first-term grades, a very satisfactory value. Except where otherwise indicated by footnotes, all correla-

Table 2
Correlations of Pre-Engineering Inventory Scores with Average First-Term Grades
(July 1944 to September 1946 Testings)

School Group	No. Cases	Test I General Verbal Ability	Test II Tech- nical Verbal Ability	Test III Compre- hension Scientific Materials	Test IV General Mathe- matical Ability	Test V Compre- hension Mechanical Principles	Test VI Spatial Visual- izing Ability	Test VII Under- standing Modern Society	Com- posite Score (II+III +IV)
A	285	.51	.55	.64	.58	.38	.42	.53	.68
B	176	.42	.52	.65	.67	.55	.35	.53	.67
C	75	.24*	.47	.53	.65	.50	.27*	.41	.67
D	403	.39	.50	.58	.63	.55	.36	***	.66
E	52	.37	.43	.54	.71	.35*	.42	.27**	.65
F	79	.50	.52	.50	.63	.40	.32	.46	.61
G	391	.35	.48	.56	.58	.35	.35	.42	.61
H	228	.39	.46	.50	.58	.42	.37	.38	.59
I	189	***	.54	.27	.46	***	***	***	.58
J	87	.16**	.31	.44	.60	.33	.36	.25*	.51
K	84	.34	.30	.45	.59	.36	.30	.42	.50
L	333	.34	.38	.45	.42	.37	.28	.39	.48
M	195	.14*	.25	.41	.51	.30	.22	.28	.44
N	100	.20*	.29	.38	.38	.22*	.16**	.33	.38
Median		.35	.48	.50	.58	.37	.35	.40	.60

* There is more than 1 chance in 100, but no more than 1 chance in 20, that a correlation as large as this could arise solely from sampling fluctuations.

** There is more than 1 chance in 20 that a correlation as large as this could arise solely from sampling fluctuations.

*** This test was not given at this school.

Table 3

Validity of Composite Score for School Groups for Which Achievement Records Covering More Than One Term Were Available
(July 1944 to November 1945 Testings)

School Group	Average Grades							
	1st-Term		2nd-Term		Two-Term		Three-Term	
	Valid-ity	No. of Cases	Valid-ity	No. of Cases	Valid-ity	No. of Cases	Valid-ity	No. of Cases
A	.68	(285)	.56	(208)	.65	(208)		
B	.67	(176)			.70	(84)		
D	.66	(403)	.43	(338)	.53	(338)		
G	.61	(391)			.50	(293)	.38	(94)
H	.59	(228)	.51	(195)	.58	(195)		
K	.50	(84)			.63	(68)	.58	(51)
L	.48	(333)			.50	(274)		
M	.44	(195)			.52	(156)		

tions in these tables are significant at the one per cent level—there is less than one chance in one hundred that correlations as large as these could arise solely from sampling fluctuations. The *Composite Score* validities are without exception significant in this sense.

Within a number of the school groups there was great variability in the course of study taken by different students—a situation that inevitably reduces the *Composite Score* validity coefficients, since the average grades do not

have a consistent meaning from student to student. Another situation that reduces the validity of the *Composite Score* in some groups is the inclusion in the grade average of such subjects as foreign languages, military science, social sciences, ethics, Bible, music, and so forth.

As would be expected, investigation reveals a tendency for school groups where such situations exist to have *Composite Score* validities below the median validity for all schools.

Table 4

Median Pre-Engineering Inventory Test Intercorrelations for Fourteen School Groups

	I	II	III	IV	V	VI	VII	Composite Score (II+III+IV)
I. General Verbal Ability		.63	.63	.38	.32	.20	.66	.60
II. Technical Verbal Ability	.63		.70	.56	.41	.32	.53	.87
III. Ability to Comprehend Scientific Materials	.63	.70		.70	.61	.51	.69	.91
IV. General Mathematical Ability	.38	.56	.70		.55	.50	.49	.84
V. Ability to Comprehend Mechanical Principles	.32	.41	.61	.55		.54	.42	.59
VI. Spatial Visualizing Ability	.20	.32	.51	.50	.54		.26	.50
VII. Understanding of Modern Society	.66	.53	.69	.49	.42	.26		.64
Composite Score (II+III+IV)	.60	.87	.91	.84	.59	.50	.64	

School groups M and N of Table 2—the two groups with the lowest validities—are both characterized by great variability in the course of study pursued by different students. School group L and school group M both consist of two or more distinct groups of students matriculating at different times of year, and this fact probably accounts in part for the low validities obtained for these groups. Group K was a summer-school group that did not take a full term's work during their first "term" of study—a fact that would tend to reduce the reliability of the average grade and thus reduce the

Composite Score validity for this group.

Of the groups with first-term validities below the median, four (H, K, L, M) have *Composite Score* validities for more than one term of work. In school group H the cumulative two-term validity is very slightly lower than the first-term validity. In the other three school groups the validities for predicting the two-term averages are higher than those for predicting the first-term averages.

2. The Test Intercorrelations. Relation-

Table 5

Comparison of Validity Coefficients Obtained for Optimally Weighted Averages of Various Tests with the Coefficients Obtained for a Single Pre-Engineering Inventory Score

School	Scores Used for Prediction	Validity Coefficients*	Optimal Weights**						
			1. General Verbal Ability	2. Technical Verbal Ability	3. Ability to Comprehend Scientific Materials	4. General Mathematical Ability	5. Ability to Comprehend Mechanical Principles	6. Spatial Visualizing Ability	7. Understanding Modern Society
A	3	.644							
A	Composite	.678							
A	2, 3, 4	.684		.14	.37	.26			
A	2, 3, 4, 7	.696		.12	.29	.24			.17
A	All tests	.699	-.01	.12	.28	.22	-.03	.09	.18
D	4	.632							
D	Composite	.659							
D	2, 3, 4	.675		.17	.16	.44			
D	All tests	.678	-.05	.19	.13	.41	.09	-.01	***
L	3	.508							
L	Composite	.534							
L	2, 3, 4	.537		.14	.28	.18			
L	All tests	.554	-.01	.17	.14	.17	-.02	.05	.19
M	4	.537							
M	Composite	.497							
M	2, 3, 4	.549		.04	.13	.42			
M	2, 3, 4, 7	.559		.03	.02	.44			.15
M	1, 4, 5	.542	.05	—	—	.57	-.08		
M	1, 4, 5, 7	.574	-.12	—	—	.55	-.15		.28
M	All tests	.584	-.19	.10	.14	.46	-.21	.03	.24

* The group of students on which Table 2 is based is not in every case exactly the same as the group used in the multiple correlation study. This fact causes a slight discrepancy between the correlations obtained in some cases.

** The weights given are multiple regression weights for the case when the standard deviations of the test scores have been equalized.

*** School D did not administer *Understanding of Modern Society*.

ships among the individual tests are shown in Table 4, which presents for each possible pair of tests the median of the fourteen separate intercorrelations obtained for the fourteen separate school groups. All the median correlations are of such magnitude that they may be taken to represent real relationships among the tests.

Multiple Correlation Study

The three tests included in the *Composite Score*—*Technical Verbal Ability*, *Ability to Comprehend Scientific Materials*, and *General Mathematical Ability*—are in general the single tests most predictive of engineering college grades. This may be seen by reference to Table 2 in which the median of the correlations of each test with first-term average grades is presented. However this is not true for every group. It can be seen that for some schools the *General Verbal Ability*, *Ability to Comprehend Mechanical Principles*, and *Understanding of Modern Society* scores had higher correlations with first-term average grades than did *Technical Verbal Ability*. This raises the question as to whether or not some alternative combination of tests would provide a better composite score for predicting average grades.

It was further found in Table 2 that in five of the fourteen school groups the correlation with first-term average grades was actually lower for *Composite Score* than for *General Mathematical Ability*. In two additional groups the correlation with average grades was at the same level for both *Composite Score* and *General Mathematical Ability*. For these groups some other weighted average of test scores would undoubtedly have provided better prediction of average grades than did the *Composite Score*.

A multiple correlation study was therefore undertaken to obtain evidence as to whether or not some other weighted average of *Pre-Engineering Inventory* scores would be preferable to the *Composite Score* now in use. Four schools, each with a large number of students, were selected for study so as to be as representative as possible of the schools for which data were available. For each school it was determined what weights should be assigned to each of the *Pre-Engineering In-*

ventory test scores in order that the weighted average of all the tests should provide the mathematically best possible prediction of average grades. The correlation of average grades with this optimally weighted average of all tests was computed as a measure of the effectiveness of the best prediction obtainable (see Table 5). This multiple correlation may be considered to be a validity coefficient.

As an example of how Table 5 may be interpreted, it is seen that in School A the optimum weights for Tests II, III, IV, and VII are in the ratio of 12:29:24:17. The corresponding weighted average of these tests correlates 0.696 with average grade, as compared to the comparable value of 0.678 for *Composite Score*.

In three of these schools it was found that the *Composite Score* functioned almost as well for predicting first-term average grade as would the optimally weighted average of all *Pre-Engineering Inventory* scores. In the fourth school the *Composite Score* did not function at all adequately from this point of view. Two facts about this school may be relevant here: first, practically none of the students at this school took English, or any related subject, during the first term; and second, the group studied in this school was composed of several entering freshman classes that had been combined in order to obtain an adequate number of cases for the statistical analysis.

The weights obtained in Table 5 apply for the case when the scores on the tests have been adjusted so that all scores have equal standard

Table 6
Optimal Relative Weights * for Raw Scores on Three
Pre-Engineering Inventory Tests for the Pur-
pose of Predicting Average Grade

School	Technical Verbal Ability	Ability to Comprehend Scientific Materials	General Mathe- matical Ability
A			1.2
D	0.5	1.3	1.7
L	0.7	0.6	1.0
M	0.7	1.3	2.1
	0.2	0.7	

* These relative weights are constructed so that their sum is 3.0, so as to facilitate comparison with the weights used to obtain the *Composite Score*, which are 1.0:1.0:1.0.

Table 7
Correlations Between Pre-Engineering Inventory Scores and Course Grades
(403 Students tested July 1944)

Tests*	First-Term Average Grade	Course Grades				
		Calculus	Chemistry	Physics	Engineering Drawing	English
I. General Verbal Ability	.39	.24	.37	.32	.15	.37
II. Technical Verbal Ability	.50	.43	.48	.43	.19	.32
III. Ability to Comprehend Scientific Materials	.58	.47	.52	.51	.31	.33
IV. General Mathematical Ability	.63	.55	.58	.58	.26	.31
V. Ability to Comprehend Mechanical Principles	.55	.44	.48	.51	.37	.26
VI. Spatial Visualizing Ability	.36	.24	.28	.30	.44	.21
Composite Score (II+III+IV)	.66	.56	.60	.59	.23	.37

* Test VII was not administered at this institution.

deviations. Table 6 presents the optimal relative weights for application to actual raw scores. The data shown in Table 6 for the four schools studied give some indication that better prediction of first-term grades might be obtained if *Technical Verbal Ability* were less heavily weighted in obtaining the *Composite Score*. The improvement of prediction so obtained would probably be of little, if any, practical importance in the case of most schools. A few of the schools, however, would probably benefit appreciably if they could construct their own composite score in accordance with their local needs. Obviously no single composite score will provide optimum prediction for every school.

Validity for Predicting Specific Course Grades⁵

The correlations between the *Pre-Engineering Inventory* scores and first-term freshman grades in individual courses have been obtained for one of the larger colleges of engineering. The records of 403 students who took an identical course of study were available for this institution. Table 7 presents the correlations of the various test scores and of the *Com-*

posite Score with grades in individual college subjects and with term averages. All correlations are statistically significant at the one per cent level.

Validity for Predicting Achievement at the Sophomore Level

The efficiency of the *Pre-Engineering Inventory* tests for predicting engineering school achievement has also been investigated by obtaining the correlations between the *Pre-Engineering Inventory* tests and the various *Engineering Achievement Tests*. The *Engineering Achievement Tests* are designed to measure achievement in several specific branches of the beginning engineering curriculum. The seven tests in this battery measure the students' knowledge of fundamental engineering principles and terminology and their application to the solution of specific problems. The *Achievement Tests* are available to colleges of engineering participating in the Measurement and Guidance Project in Engineering Education and are used primarily in the examination of regularly enrolled sophomore students at or near the end of the academic year.

Data were available for 430 students in one school who took the *Pre-Engineering Inventory* in September 1946 and the *Engineering Achieve-*

⁵ This study was planned by and carried through under the direction of K. W. Vaughn as Director of the Graduate Record Office.

Table 8

Correlations Between Pre-Engineering Inventory Scaled Scores and Engineering Achievement Test
(Form B) Raw Scores
(236 Students at School Q Tested in October 1946 and May 1948, and 430 Students
at School R Tested in September 1946 and May 1948)

Pre-Engineering Inventory Tests	School	Engineering Achievement Tests						
		English Expres- sion	Engi- neering Drawing	General Chem- istry	Mathe- matics	Physics (Mechan- ics)	Elec- tricity and Mag- netism	Heat, Light, and Sound
I. General Verbal	Q	.70	.32	.46	.35	.38	.39	.40
Ability	R	.58	.36	.33	.23	.27	*	*
II. Technical Verbal	Q	.52	.41	.66	.30	.43	.44	.48
Ability	R	.37	.42	.48	.30	.35	*	*
III. Ability to Comprehend	Q	.58	.52	.59	.46	.55	.54	.53
Scientific Materials	R	.51	.47	.49	.43	.57	*	*
IV. General Mathematical	Q	.52	.35	.51	.44	.55	.40	.42
Ability	R	.39	.37	.41	.51	.51	*	*
V. Ability to Comprehend	Q	.43	.54	.49	.47	.54	.46	.53
Mechanical Principles	R	.30	.46	.38	.39	.53	*	*
VI. Spatial Visualizing	Q	.36	.52	.33	.36	.38	.26	.33
Ability	R	.18	.32	.14	.21	.26	*	*
VII. Understanding of	Q	.51	.32	.46	.34	.43	.41	.44
Modern Society	R	.48	.33	.37	.24	.39	*	*
Composite Score	Q	.62	.49	.67	.46	.58	.53	.55
(II+III+IV)	R	.51	.50	.54	.50	.57	*	*

* Not administered.

ment Tests in May 1948, and for 236 students in a second school who took the *Inventory* in October 1946 and the *Achievement Tests* in May 1948. The correlations appear in Table 8.

On the whole the results are quite good in view of the fact that a lapse of time as great as two years intervenes between the two testings. Even the lowest correlations in this table are significant at the one per cent level. The most striking thing about the data presented in Table 8 is that the *Composite Score* is the best or second best predictor of achievement, as measured by the *Engineering Achievement Tests*, in almost all of the several areas of engineering education for both school groups. In those cases in which the *Composite Score* is not the best predictor, one of the tests comprising the *Composite Score* is the best predictor of achievement, with few exceptions.

This study further substantiates findings that the *Pre-Engineering Inventory* in general and the *Composite Score* in particular are valid predictors of engineering school success.

General Conclusions

The results of the several test analyses described above indicate that the *Pre-Engineering Inventory Composite Score* on the whole performs very satisfactorily as a predictor of engineering school success. Of course, any fixed set of weights will not provide optimum prediction for every school. The use of a specially weighted combination of those tests that comprise the *Composite Score*, or possibly of some other tests, would undoubtedly appreciably improve the prediction in some schools. How-

ever, it is not certain that the advantages gained by computing a special composite score would justify the expense involved.

An examination of the correlations presented here for the *Pre-Engineering Inventory* tests not included in the *Composite Score* is of value for understanding these tests and their relation to the engineering school curriculum. The utility of these tests cannot be unequiv-

ocally evaluated by the approach used in the studies covered here, since these tests have functions beyond the prediction of college grades, particularly as aids in the guidance of the student and in the selection of broadly gifted individuals who will do credit to the profession after graduation.

Received November 18, 1949.
Early publication.

The Kuder Literary Scale as Related to Achievement in College English *

A. Kimball Romney

University of Wisconsin

The relation between scores obtained on different types of interest inventories and achievement has been summarized and appraised elsewhere.¹ Most of the studies which examined the relation between Kuder interests and achievement showed positive relationships.² It can be added that the studies were done in the main with rather small groups using, most frequently, college grades as a criterion of achievement. Whether results from larger groups using more refined methods would give different results has not been known.

The present note is concerned with reporting results obtained using a precise index of achievement on a fairly homogeneous group of over a thousand subjects.

Specifically it concerns correlation data between the Kuder literary score, ACE scores (American Council on Education Psychological Examination for College Freshmen, 1947 edition), and achievement in college English classes. It is the nature of the achievement score that makes these data of special interest. They are noteworthy for the following reasons: (a) the group of subjects is composed of 1085 (566 male, 519 female) freshman students who took English 1 Autumn Quarter at Brigham Young University in 1947; (b) since it included all new freshmen the results are somewhat free of many errors that might have been introduced by sampling procedures; (c) the subjects were all exposed to as near the same amount of instruction as possible during the quarter; (d) the achievement is judged not on the basis of a grade (as "A", "B", etc.) but rather on the basis of a long (554 item), objective, carefully administered achievement test given at the

end of the fall quarter; and (e) college aptitude as measured by the ACE is taken into account as an important variable.

The English Achievement Examination

The test, which covered all the areas taught in English 1 during Autumn Quarter 1947, was prepared by a faculty committee before the beginning of the quarter. It consisted of 554 items arranged into five sections: vocabulary, essays, short stories, grammar, and miscellaneous. At the beginning of the quarter all English teachers received a letter from the department head indicating the areas of the test, amount of time which would be allowed for each area, and when it would be given. This information was given to the students at the same time. Similar letters were sent out at mid-term and three weeks before the end of the quarter. No teacher knew the exact questions and none of the teachers who taught in the experiment had seen the test before it was administered.

At the beginning of the quarter the students were divided into ability areas on the basis of the Purdue English Placement Test. Teachers were assigned to teach the respective groups, i.e., each taught some groups classified as "poor," some classified as "intermediate," and some groups classified as "good." This scattered the teachers over the entire classification of groups equalizing the teaching factor as nearly as possible. All teachers used the same text and the same units of work for all classes.

At the close of the quarter, qualified teachers from the university were called in by the English department and were given instructions as to the administration of the achievement test and it was administered to all students the same afternoon under the same conditions. The teachers and tutors who administered it had all received special training for that purpose and the test, which was under

* The author is grateful to Dr. Antone K. Romney, Head of the Counseling Service at Brigham Young University, for permission to collect the data reported here. The results do not necessarily reflect the policy or the nature of the work done by the Counseling Service.

¹ Super, D. E. *Appraising vocational fitness*. New York: Harper and Brothers. 1949. Pp. 727.

² Super, D. E. *Op. cit.*, pp. 457-458.

the supervision of the Counseling Service, was scored on International Business Machine equipment.

Results

The various correlation coefficients obtained are shown in Table 1.

It can be seen that the Kuder literary scale has a small but statistically significant correlation with English achievement which is roughly .3 for both male and female. This does not rise significantly when possible effects of ACE are partialled out.

The correlation between ACE and English achievement was high for both males and females, .69 and .84 respectively.³

With reference to the multiple R, it can be seen that no significant changes result by the addition of the Kuder literary scale. It is to be noted that the Kuder scale and the ACE are apparently quite independent of each other.

It can be concluded that as far as these data are concerned the correlation between

Table 1

Correlation Coefficients between English Achievement, Kuder Literary Scale, and ACE

Dependent Variable	Independent Variables			
	Kuder	ACE	Kuder Plus ACE (a)	Kuder with ACE (b)
English Achievement Male (N = 566)	.27±.04	.69±.02	.71±.02	.28±.04
English Achievement Female (N = 519)	.29±.04	.84±.01	.85±.01	.29±.04

(a) Multiple correlation.

(b) Partial correlation.

achievement in a college English class and the Kuder literary scale is very low, even though statistically significant.

Received October 5, 1949.

Early publication.

³ No explanation is offered for the difference in coefficients observed between male and female scores.

Scores on the Strong Vocational Interest Blank and the Kuder Preference Record in Relation to Self Ratings

Ralph F. Berdie

Student Counseling Bureau, Office of the Dean of Students, University of Minnesota

Expressed interests and measured interests are far from identical. In terms of common elements, perhaps no more than between 25 and 50 per cent of the factors associated with one are associated with the other. Various studies report correlation coefficients ranging about .50 between measured and expressed interests, depending upon the test used, the method employed for eliciting and classifying expressed interests and the sample.

In a study of the Strong Vocational Interest Blank scores of 1000 men who came to a University Testing Bureau, Darley (5) reported contingency coefficients between claimed vocational choices and interest score patterns ranging from .35 to .57. He concluded, "Claimed vocational choices cannot be substituted for measured interests in effective counseling."

Lalegar (11) also used the Strong Blank, along with the Manson Occupational Blank, to determine the relationship between the test scores and the occupational choices of 703 eleventh grade girls, an age group for which the Strong Blank is not too appropriate. She concluded: "Insofar as stated choice of occupation by groups of individuals may be considered a true criterion of interest, the lack of relationship between the statement of occupational choice and interest scores or letter ratings may be considered evidence of the lack of validity of the interest inventory." As one must disagree with Lalegar's assumption regarding the "true criterion" of interests, so must one also reject her conclusion that vocational interest inventory scores cannot contribute materially to vocational planning.

Kopp and Tussing (9) report the relationship between scores on the Kuder Preference Record and responses to a simple questionnaire used for appraising vocational interests. For each student, the nine scores on the Kuder Record were ranked in order from one to nine,

and these rankings were correlated with the orders in which the occupations were ranked by the students. The authors do not report how these correlation coefficients were obtained, but they state that for 115 boys in grade 10B, the correlation between questionnaires and Kuder scores was .59 and for 117 girls in grade 10B, the correlation was .50.

Rose (14) gave the Kuder blank to 60 "unselected" veterans referred to a Vocational Advisement Unit. Each veteran was then given nine cards. On each card were 17 occupations selected from Kuder's manual (10) as characteristic of a given area. The subjects ranked the nine areas according to preference for the occupation. The coefficient of contingency obtained was .61. The rank order correlation coefficients for each of the 60 cases ranged from -.05 to .99, with a median r_{ho} of .64.

Crosby and Winsor (4) administered the Kuder Preference Record to 222 men and women sophomores in a college of agriculture and home economics. These students, after an explanation of the type of interests measured by the test, were asked to estimate their percentile scores, using the Kuder profile sheets. The correlation coefficients between obtained scores and estimated scores on the seven scales ranged from .39 to .66. The median r was .54.

Information is available about the relationship between the Kuder and the Strong tests (18), but no study reported to date allows direct comparisons between the relationship of the Kuder Record and self ratings and the Strong Blank and self ratings. Paterson (13), reporting the case of a returning veteran who took both tests, raises a question concerning the relative ease of "faking" the tests. He suggests: "The Kuder and Strong inventories both yield important information about an individual's interest patterns, when obtained

in a guidance situation. However, in a selection situation, it would appear that the Strong is to be preferred because it is more subtle and the vocational significance of liking or disliking each of the 400 items is not so readily apparent to the person taking the test."

Longstaff (12) reports an investigation of the fakability of the Strong test and the Kuder test where 35 men and 24 women took the Strong test, men's form, and 37 men and 22 women took the Kuder test. The subjects first took the tests under "normal" conditions and then took them with directions to raise their scores on certain scales and lower them on others. The results indicated both tests could be faked and that some scales were more fakable than others. The Strong test was easier to fake upwards, the Kuder test easier to fake downward.

A theory of vocational interests with particular reference to the Strong Blank has been

Kuder Preference Record and self ratings, using the same sample. Does the picture a person has of his own interests correspond more closely to the picture given by the Strong Blank or to that given by the Kuder Record?

Method

Each man who came to the Student Counseling Bureau of the University of Minnesota during the first part of 1948 and who was to take an interest test was given the Strong test, the Kuder test and a self rating form.

The rating form was of the graphic rating type and covered the following nine occupational areas: (1) biological sciences; (2) artistic creation and appreciation; (3) physical sciences; (4) technical occupations; (5) social service; (6) musical occupations; (7) business detail; (8) selling; and (9) verbal or literary. For each area the graphic scale was identical. The following is an example:

1. This occupational area centers about the <i>biological sciences</i> and includes medicine, dentistry and psychology.				
My interests are very much unlike interests of people in this area	Somewhat dissimilar	No marked similarity or dissimilarity	Somewhat similar	My interests strongly resemble the interests of people in this area

discussed by Bordin (2). He presents the hypothesis that, "In answering a Strong Vocational Interest Test, an individual is expressing his acceptance of a particular view or concept of himself in terms of occupational stereotypes." In support of this hypothesis, Bordin cites the relationship between claimed and measured interests and also the ability of subjects to manipulate or fake their vocational interest patterns. These phenomena can be explained making use of concepts other than those used by Bordin and we already have suggested here, in reference to the relationship between measured and claimed interests, the effect of common elements.

Purpose

The purpose of this investigation was to determine the relative agreement between scores on the Strong Vocational Interest Blank and self ratings and between scores on the

The order of presentation of the tests and the rating form was varied systematically so that one-third of the 500 men tested took the Strong test first, one-third took it as the second test and one-third took it as the last test in the series. Similarly, the order of presentation for the other test and rating form was rotated.

The ages of the 500 men varied from 14 years through 37 years, with a mean age of 20.8 years, a median age of 20.6 years and a standard deviation of 3.5 years. Only seven men were 16 years of age or younger, only 11 were 30 years old or older.

The largest single group, the pre-college group (N=195), consisted of people not yet registered within the University but planning on matriculating within one calendar year. The largest proportion of these people were completing or just had completed their senior year in high school. The non-college group (N=19) consisted of people who were not in residence in the University and who were not

planning on entering within the next year. The students in the College of Science, Literature and the Arts ($N=126$), the Institute of Technology ($N=89$), the College of Agriculture, Forestry and Home Economics ($N=13$), the College of Education ($N=3$), and the College of Pharmacy ($N=2$) ranged from

students beginning their first year of college to students completing their fourth year. Similarly, students from the remaining colleges tended to come from all classes within those colleges. Most of these 500 students, however, were people who had completed less than two years of college.

Table 1

Categories Used in Classifying Scales of the Strong Blank, the Kuder Record and the Self-Rating Form with the Number of Subjects Having Different Types of Patterns or Scores in Each Area on Each Test

Strong (artist, architect)		Artistic Self-Rating (item 2)		Kuder (artistic scale)	
Type of Pattern	Number	Score	Number	Score	Number
primary pattern	13	5	55	75-100	120
secondary pattern	15	4	136	65-74	48
tertiary pattern	46	3	136	50-64	57
no pattern	426	2	91	0-49	275
		1	82		
Strong (osteopath, physician, psychologist, dentist)		Scientific—Biological Self-Rating (item 1)		Kuder (scientific scale)	
Type of Pattern	Number	Score	Number	Score	Number
primary pattern	47	5	79	75-100	114
secondary pattern	38	4	193	65-74	46
tertiary pattern	60	3	98	50-64	77
no pattern	355	2	67	0-49	263
		1	63		
Strong (mathematician, physicist, chemist, engineer)		Scientific—Physical Self-Rating (item 3)		Kuder (scientific scale)	
Type of Pattern	Number	Score	Number	Score	Number
primary pattern	27	5	127	75-100	114
secondary pattern	7	4	161	65-74	46
tertiary pattern	26	3	72	50-64	77
no pattern	440	2	71	0-49	263
		1	69		
Strong (Farmer, aviator, carpenter, printer, M & P.S. teacher, policeman, forest service man)		Technical Self-Rating (item 4)		Kuder (mechanical scale)	
Type of Pattern	Number	Score	Number	Score	Number
primary pattern	150	5	98	75-100	97
secondary pattern	72	4	169	65-74	47
tertiary pattern	95	3	111	50-64	52
no pattern	183	2	62	0-49	304
		1	60		

Table 1 (Continued)

Strong (YMCA phys. dir., personnel dir., pub. admin., YMCA secy., social science teacher, school supt., minister)		Social Service		Kuder (social service scale)	
Type of Pattern	Number	Score	Number	Score	Number
primary pattern	71	5	91	75-100	246
secondary pattern	57	4	173	65-74	40
tertiary pattern	83	3	100	50-64	63
no pattern	289	2	73	0-49	151
		1	63		

Strong (Musician)		Musical Self-Rating (item 6)		Kuder (musical scale)	
Type of Pattern	Number	Score	Number	Score	Number
primary pattern	74	5	51	75-100	175
secondary pattern	69	4	109	65-74	53
tertiary pattern	77	3	104	50-64	66
no pattern	280	2	95	0-49	206
		1	141		

Strong (CPA, accountant, office mgr., purch. agent, banker, mortician)		Business Office (Clerical) Self-Rating (item 7)		Kuder (Computational and Clerical Scales)		
Type of Pattern	Number	Score	Number	Score	Number (computational)	Number (clerical)
primary pattern	81	5	70	75-100	151	112
secondary pattern	78	4	152	65-74	26	29
tertiary pattern	96	3	118	50-64	55	56
no pattern	245	2	83	0-49	268	303
		1	77			

Strong (Sales manager, real estate sales, life insurance sales)		Sales Self-Rating (item 8)		Kuder (Persuasive Scale)	
Type of Pattern	Number	Score	Number	Score	Number
primary pattern	128	5	64	75-100	271
secondary pattern	94	4	156	65-74	46
tertiary pattern	89	3	123	50-64	37
no pattern	189	2	64	0-49	146
		1	93		

Strong (Advertising, lawyer, author journalist)		Verbal—Literary Self-Rating (item 9)		Kuder (Literary Scale)	
Type of Pattern	Number	Score	Number	Score	Number
primary pattern	48	5	77	75-100	177
secondary pattern	51	4	157	65-74	55
tertiary pattern	140	3	111	50-64	57
no pattern	261	2	87	0-49	211
		1	68		

The 37 occupational scales of the Strong Blank were divided into nine groups, as shown in Table 1, and the 9 scales of the Kuder test were similarly classified. Two scales of the Kuder, computational and clerical, were classified as Business-Office, and the Kuder Scientific scale was considered as being both "Biological science" and "Physical science."

Thus, for each student were available 37 scores on the Strong, nine scores on the Kuder and nine scores on the self rating form. A pattern analysis (5) was made of the Strong scores, according to the groups in Table 1. If within a group, a plurality of the scores were A's and B+'s, that was a primary pattern. If a plurality of scores were B+'s and B's, that was a secondary pattern. If a plurality of scores were B's and B-'s, that was a tertiary pattern. All other groups were labeled "no pattern." In case of the musician's scale, an A was called a primary pattern, a B+, a secondary pattern, and a B, a tertiary pattern.

A percentile score of 75 through 100 on the Kuder was called a primary pattern, a percentile score of between 65, through 74, a secondary pattern and a percentile score of between 50 through 64 a tertiary pattern. All other scores were "no pattern." The self ratings in each area were given values of from one to five, with five indicating greatest similarity of interests.

Contingency coefficients and chi squares were then computed to indicate the degree and significance of relationship between the Strong Blank and self ratings and between the Kuder Test and self ratings.¹

Results

The three indices of interests first can be compared on the basis of the frequency of significant scores in the nine areas. Kuder says that percentile scores of 75 and above on his test are significant for purposes of vocational counseling (10). Strong says that scores of A and B+ on his test are significant (16), and Darley states that primary patterns on the Strong test are most significant (5).

Table 1 reveals that the number of significant scores, as defined by these authorities, is

¹ The formula for the contingency coefficient was obtained from Guilford, J. P., *Psychometric methods*, New York: McGraw Hill Book Co., 1936, pp. 357-359.

much greater on the Kuder test in eight of nine areas and in only one area, the sub-professional or technical, are there more significant scores on the Strong test.

On the Kuder test, 24 per cent of the group had significantly high scores on the artistic scale, while on the Strong, only three per cent had significantly high scores. Only 15 per cent had *any* type of interest pattern in this area on the Strong. The distribution of self ratings was more symmetrically distributed, with 11 per cent claiming much interest in this area.

Distributions in the musical area are somewhat similar to those in the art area, with 15 per cent obtaining A's or B pluses on the musician's scale of the Strong and 35 per cent obtaining percentile scores of 75 or more on the musical scale of the Kuder. On the self ratings, many more men indicated no interest in this area than there were men who indicated much interest.

Again in the scientific areas, relatively few significant scores were found on the Strong profile, while many were found on the Kuder. Combining the biological and physical science groups on the Strong, only 74 primary patterns were identified. On the Kuder scientific scale, 23 per cent received percentile scores of 75 or above. The differences in the social service area are even greater, with 14 per cent having primary patterns on the Strong and 49 per cent having high scores on the Kuder. The trends are similar in the business and literary areas.

In the technical and mechanical area, 30 per cent had primary interest patterns on the Strong and only 19 per cent had high scores on the Kuder. In only one of the nine areas, the sales area, were there more primary Strong patterns than there were in this technical area. On the Kuder, fewer people had scores of 75 or more on the mechanical scale than on any other scale.

Thus, on the Kuder, the areas having the greatest number of high scores were the sales (persuasive) and the social service areas and on the Strong, the areas having most primary patterns were the sales and the technical areas. The areas on the Kuder test having the fewest high scores were the mechanical and clerical areas and on the Strong the areas having the

fewest primary patterns were the artistic and the physical science areas.

In terms of self ratings, most students rated themselves as very much interested in physical science and technical work and fewest rated themselves as very interested in music and art.

On the Strong test there was a total of 639 primary patterns, or an average of 1.3 per student and a total of 481 secondary patterns. Combining these figures, the average student had 2.2 primary and secondary patterns. On the Kuder, 1463 scores were at or above the 75th percentile. The average student had 2.9 scores which according to Kuder are significant in counseling. In the norm group upon which these percentile scores were based, for a group of 500 men, one would find 1125 percentile scores of 75 or above out of the 4500 possible scores available on the nine scales of the test or an average of 2.3 significant scores per man.

The relationship found here between measured and expressed interests approximates that reported in previous studies. The median contingency coefficient between the Strong Test and self ratings was .43, between the Kuder test and self ratings, .52. Table 2 presents the coefficients showing the degree of relationship in each area between each of the

Table 2

Contingency Coefficients Showing Relationships Between Self Ratings of Vocational Interests and Scores on the Strong Vocational Interest Blank and on the Kuder Preference Record

Occupational Area	C with Strong	C with Kuder
Technical	.55	.47
Computational	.61	.34
Physical Sciences	.32	.46
Social Service	.43	.52
Musical	.39	.60
Sales	.58	.58
Biological Sciences	.27	.30
Verbal—Literary	.51	.61
Artistic	.33	.58
Clerical	(.61)*	.52

* This is the same statistic as for "Computational."

two tests and self ratings. The *chi* squares were all statistically significant beyond the one per cent level of probability.

In two areas, the Strong scores are more closely related to self ratings than are the Kuder scores. In five areas, the Kuder scores are more closely related to self ratings. The range of contingency coefficients for the Strong is from .27 to .61, for the Kuder, from .30 to .61.

On both tests, self ratings of interest in the biological sciences are least related to relevant scores on the tests. The next poorest agreement is found in the physical science area where self ratings have a relationship in terms of contingency coefficients to Kuder and Strong patterns of .46 and .32 respectively. Self ratings in the sales area tend to be related to patterns on both tests.

Disagreement is found in the clerical and computational areas. The patterns on the Business-Office group of the Strong tend to be accompanied by similar self ratings. The Kuder computational scores were only slightly related to these self ratings. The Kuder clerical scores were more related to clerical self ratings than were Kuder computational scores.

In the case of those areas which have strong avocational implications, art and music, the relations between self ratings and Kuder scores are greater than those between self ratings and Strong scores. The explanation is perhaps that most students are unable to differentiate their thinking about vocational and avocational interests. Clinical experience suggests these two scales of the Kuder test are much more measures of avocational interests than of vocational interests.

When the results obtained here are compared to those reported by Crosby (4), the relative ease of self-estimating scores in the persuasive area is apparent. Crosby found, for two groups, correlations of .62 and .66 between Kuder persuasive scores and self-estimated scores in this area, as compared to a contingency coefficient of .58 found here. Similarly, in the literary area, Crosby's correlations of .51 and .61 can be compared to the contingency coefficient of .61 obtained here between self ratings and the Kuder.

Discussion

The results presented here are in general agreement with the results obtained by other investigators and the correlation between measured and self-estimated interests approximates .50. In agreement with Paterson's hypothesis concerning the relative subtlety of the two tests, scores on the Kuder tend to have a closer relationship to self ratings of interests than do the scores on the Strong. This may be a function not only of the items in the tests but also of the categories used in grouping the scales and defining the self ratings, although these categories were achieved through careful study of both tests.

The men studied here found it relatively difficult to estimate their own scientific interests, as they were measured by the tests, and on the other hand, the men were more able to estimate their persuasive and sales interests. In no occupational area, however, was there close enough agreement between measured interests and self-estimated interests to suggest that counseling can be done on the basis of one or the other. As long as measured interests have a relevancy for vocational satisfaction and as long as self-estimated interests play an important role in the vocational deliberations of individuals, both types of interests must be considered.

The counselor working with individuals similar to those in this sample can expect to find many more significant scores, as defined by the test authors, on the Kuder blank than on the Strong test. Here each person averaged 2.9 scores of 75 or above on the Kuder and only 1.3 primary patterns on the Strong. This is a result of the Kuder norms being based on groups very similar to those studied here in contrast to the Strong scores which are based on norms derived from occupational groups (6). Both kinds of scores can be useful in counseling but the counselor should recognize that in every case but that of the technical-mechanical area, a "significant" score on the Strong test is a more rare occurrence than a "significant" score on the Kuder.

Whereas one-half of a group similar to this will have high social service scores on the Kuder, only 14 per cent will actually have interests that resemble those of men in social

service work, as implied by their primary patterns in group V on the Strong test. Over one-half, however, will indicate they bear a marked resemblance to the men in this field, as shown by their self ratings. Strangely enough, all three indices can be correct. Strong discusses the predominant similarities of interests and would perhaps agree that most of these subjects are more like men in social service than they are unlike them. The Strong blank does not take into account these similarities, however, but rather, utilizes the differences between groups, and consequently, relatively few of this sample obtain primary patterns in the social service area. In other words, the subjects who receive primary patterns in group V share with the men in the standardization groups those interests which tend to make them *different* from men in general. We can only say that those subjects who obtain high scores on the social service scale of the Kuder are showing interests which are shown by people in the social service areas, many of these interests also being held in common with men not in these areas.

One explanation for the relatively greater ease of estimating scores on the Kuder may thus be derived. In estimating Kuder scores, the subject needs consider only his similarities to men in the defined groups, but in estimating Strong scores, he needs to consider both how he resembles men in the defined group and also how he differs from men in general.

These considerations do not adequately explain the degree of agreement between self-estimates of interests and test scores. Recognizing that this agreement is reduced by the unreliability of test scores and the inconstancy of self ratings, what are those factors which are related to both interest scores and self-estimates?

Preliminary determinations of some of these factors have been reported (1). In an intensive study of 136 college students, college aptitude was found to be significantly related to measured interests in engineering but not to expressed interests. Those students with measured engineering interests participated in fewer religious activities than did those students with no measured interests. This difference in activities was not found when the group was analyzed on the basis of expressed

engineering interests. Preference for high school mathematics teachers was related to measured interests in engineering but not to expressed interests. Family background was related equally to both types of interests. When business interest groups were compared, morale scores on the Minnesota Personality Scale were related to expressed interests in business but not to measured interests, and frequency of "social" activities bore the same relationships. An elaboration of this type of study would throw light on the differences between measured and expressed interests.

Other questions arise. Why is there a tendency for scores and self-estimates in some areas to be more closely related than in other areas? What distinguishes those people who can quite accurately estimate their interests, as measured by tests, from those who cannot? Many questions remain to be answered concerning the relationships between measured and estimated interests.

References

1. Berdie, R. F. Factors associated with vocational interests. *J. educ. Psychol.*, 1943, 34, 257-277.
2. Bordin, E. S. A theory of vocational interests as dynamic phenomena. *Educ. & Psychol. Meas.*, 1943, 3, 49-65.
3. Christensen, T. E. Some observations with respect to the Kuder Preference Record. *J. educ. Res.*, 1946, 40, 96-107.
4. Crosby, R. C., and Winsor, A. L. The validity of student estimates of their interests. *J. appl. Psychol.*, 1941, 25, 408-414.
5. Darley, J. G. *Clinical aspects and interpretation of the Strong Vocational Interest Blank*. New York: Psychological Corporation, 1941.
6. Diamond, S. The interpretation of interest profiles. *J. appl. Psychol.*, 1948, 32, 512-520.
7. Gordon, H. C., and Herkness, W. W. Do vocational questionnaires yield consistent results? *Occupations*, 1942, 20, 424-429.
8. Kelso, D. F., and Bordin, E. S. The ability to manipulate occupational stereotypes inherent in the Strong Vocational Interest Test. *Amer. Psychol.*, 1948, 3, 352-353.
9. Kopp, T., and Tussing, L. The vocational choices of high school students as related to scores on vocational interest inventories. *Occupations*, 1947, 25, 334-339.
10. Kuder, G. F. *Kuder Preference Record, Revised Manual*. Chicago: Science Research Associates, 1946.
11. Laleger, G. E. *Vocational interests of high school girls*. New York: Bureau of Publications, Teachers College, Columbia Univ., 1942.
12. Longstaff, H. P. Fakability of the Strong Vocational Interest Blank and the Kuder Preference Record. *J. appl. Psychol.*, 1948, 32, 360-369.
13. Paterson, D. G. Vocational interest inventories in selection. *Occupations*, 1946, 25, 152-153.
14. Rose, W. A comparison of relative interest in occupational groupings and activity interests as measured by the Kuder Preference Record. *Occupations*, 1948, 26, 302-307.
15. Steffle, B. The reading difficulty of interest inventories. *Occupations*, 1947, 26, 95-96.
16. Strong, E. K. *Vocational interests of men and women*. Stanford: Stanford Univ. Press, 1943.
17. Thorndike, E. L. *Adult interests*. New York: Macmillan, 1935.
18. Wittenborn, J. R., Triggs, F. O., and Feder, D. D. A comparison of interest measurement by the Kuder Preference Record and the Strong Vocational Interest Blanks for men and women. *Educ. & Psychol. Meas.*, 1943, 3, 239-257.

Visual Differentiation of Moving Objects

Newell C. Kephart and Guy G. Besnard

Division of Applied Psychology, Purdue University

Much work has been done at the Occupational Research Center, Purdue University, in cooperation with various industrial plants in an attempt to find skills which are pertinent to performance on various industrial jobs. At the present time employees of cooperating plants are tested and a statistical comparison of job performance and test scores is made. Recently, research has been attempted to see whether it would be possible to attack the problem from another angle by setting up in the laboratory counterparts of industrial jobs or parts of jobs. As one such approach the study of the visual differentiation of moving objects was selected.

Experimental Procedure

Description of Apparatus and Material. The apparatus consisted of an inclined trough down which clear glass spheres could be rolled. This trough was made of wood, 46 inches long and 1.5 inches wide. The inside was lined with white paper since preliminary experiments showed that a dark lining made the task too easy. A second trough underneath the first one and sloping in the opposite direction allowed the spheres to return to the experimenter in the original sequence (see Figure 1).

Twenty clear glass marbles of the ordinary commercial type were provided. Of these, 10 were left plain and the remaining 10 were scratched with a sharp instrument so that three circles were on three circumferences in three perpendicular planes. For the purpose of terminology in the test situation the unmarked spheres were called "good" and the etched spheres "bad."

A head rest was provided to keep the subject's head in a specified position during the experiment, so that each subject would view the spheres from approximately the same distance and at approximately the same angle. In the early experiments, the spheres were released at the top of the trough and allowed to roll down by themselves. Since the spheres were immobile at the time of release, their rate of speed near the top of the trough was so slow that they were very easily seen by the subject and the discrimination was too easy. In order to make the task more difficult, we had to eliminate the possibility of seeing the sphere until it had attained adequate speed. In order to do this, a screen was placed in such a position as to permit the subject to view only the last 18 inches of the trough, but not to view the first 28 inches.

The room in which the experiment was conducted was kept in darkness except for a 10.5 inch, 40 watt tubular Lumiline bulb, mounted in a clear tin reflector. This light was placed 12.5 inches above the table and directly above the last 18 inches of the trough. By keeping the different physical aspects of the material and the apparatus constant we hoped to reduce our test to one variable, that of discriminating between the "good" and the "bad" moving spheres.

Visual skills were measured with the Bausch & Lomb Ortho-Rater. This instrument provides measures of 12 visual skills which have been found on the basis of scientific investigation to be important to success on individual jobs. The 12 skills measured are: 1. Far Phoria Vertical; 2. Far Phoria Lateral; 3. Far Acuity Both Eyes; 4. Far Acuity Right Eye; 5. Far Acuity Left Eye; 6. Depth Perception; 7. Color Vision; 8. Near Acuity Both Eyes; 9. Near Acuity Right Eye; 10. Near Acuity Left Eye; 11. Near Phoria Vertical; and 12. Near Phoria Lateral.

First Experiment. Each subject was given the rolling sphere test from two views, one from the side of the trough (where he could see the spheres rolling from right to left in front of him) and one from above the end of the trough (where he could see the spheres rolling toward him). These two trials were given at different times, usually on succeeding days.

About half of the subjects were given the "side view" first (hereafter referred to as "side view first"), and later were given the "end view" (hereafter referred to as "end view second"); the other half were given the end view (end view first) and later were given the side view (side view second). When each subject came for the first time, he was given the following verbal instructions about the experiment:

"In this experiment spheres will be rolled down the inclined trough. Some of the spheres are unmarked such as this (show example); these spheres we call 'good.' Other spheres are marked such as this (show example); these spheres we call 'bad.' Please place your head against the head rest. I will now show you an example of a 'good' sphere rolling by (show example) and here is the 'bad' sphere (show example). As each sphere rolls by you on the trough, say 'good' if you think it is a good sphere, or 'bad' if you think it is a bad one. In case you are not able to discriminate between good or bad, say, 'don't know,' but if you have any impression at all, say 'good' or 'bad,' depending upon which you think it is. It is particularly important that you make a response for each sphere. The spheres will roll by rapidly but at any given time you will see only one. Do you have any questions? . . . Here comes the first sphere!"

The 20 spheres, previously arranged in random order

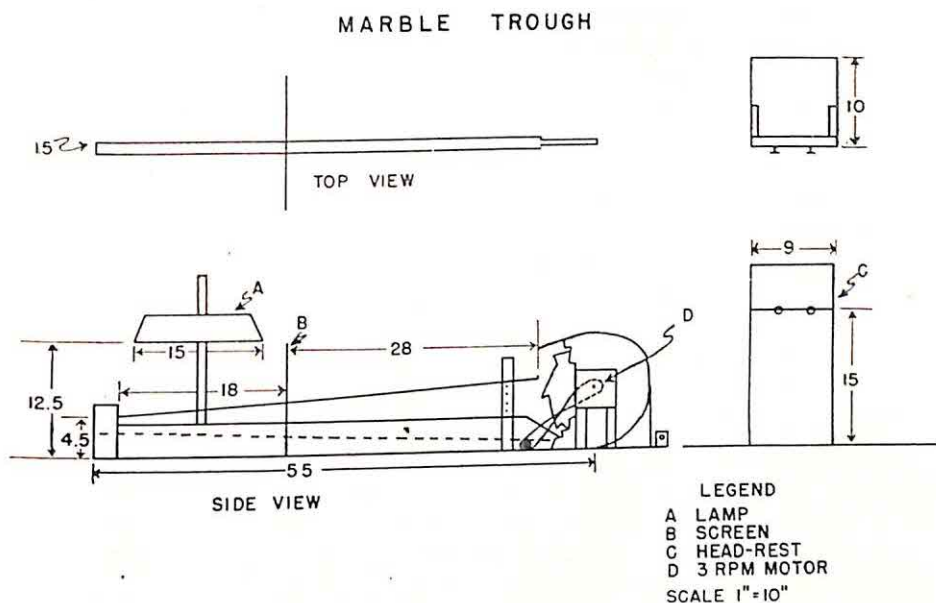


Fig. 1. Top and side views of apparatus.

and placed in the return trough, were rolled by one of the experimenters, one at a time for 8 sequences, making a total of 160 spheres for each of the two views. Each sphere was released at the top of the trough at about the time the previous sphere was passing under the screen and took approximately 2.5 seconds to roll by the entire length of the trough, but was visible to the subject for only .6 of a second.

Forty-six members of a class in beginning psychology at Purdue University were used in the first experiment. Ortho-Rater test scores were available for these students. Subjects were selected on the basis of their visual profile so that they would be fairly homogenous with regard to their near visual acuity and with as wide a range of phoria test scores as possible.

Second Experiment. The essential differences between the first and the second experiment were: 1. the addition of a motor to bring up the spheres and drop them in the trough automatically. The motor was geared to drop a sphere in the trough every 2.5 seconds. It took these spheres approximately 1 second to roll down the trough, but it was visible to the subject only about .4 second. 2. The subjects in this last experiment were not selected on the basis of their visual scores but were a random sample of a class of approximately 150. 3. The number of spheres was held to 4 sequences of 20 spheres or 80 spheres altogether, since the first experiment had shown this to be a sufficient number of trials for reliability.

The rest of the test procedure remained constant. Forty-five subjects from a beginning course in psychology were used in this experiment.

Results

The number of errors made by each subject (an incorrect response or a response of "don't

know" were considered as errors) was computed. These computations were made for each view separately by orders (first or second). In addition to the total number of errors made by each individual there are also totals for each of the sequences of 20 spheres viewed by the subject.

Reliability. Reliability was computed separately for the end and for the side view, but no differentiation was made, as to whether the view was the first or the second. Reliability was measured by the odd-even technique, taking as the "Odds" the odd numbered spheres of the odd numbered 20 sphere sequences plus the even numbered spheres of the even numbered sequences, and as "Evens" the even numbered spheres of the odd sequences plus the odd numbered spheres of the even sequences. Thus each sphere was included alternately in the "even" and "odd" for each view, since minor unintentional differences in the etching of some of the marbles, or inherent flaws, might affect the appearance of individual spheres and therefore the scores of the subject. The resulting "odd-even" corrected correlation for the side view of the first experiment was $r = .87$. The corrected reliability of the end view was $r = .95$. The reliability for the side view was influenced significantly by one exceptional case. With this

case omitted, the corrected reliability was reduced to $r=.64$. The apparently lower reliability of the side trial probably is, in large measure, due to the markedly lower variability of scores for the side due to the low number of errors.

The reliability as shown in the second experiment was measured in the same manner as in the first experiment. The corrected coefficient for the end view, in the second experiment, was .89. The corrected reliability of the side view for the second experiment was .89.

It will be noted that in the second experiment the reliability of both the side and the end view are approximately the same. The higher reliability from the side view in the second experiment can probably be explained by the fact that the variability of scores from the side view was larger due to increased difficulty of the test.

Differences Between End and Side Views. The mean number of misses made by each group on the end and on the side was computed. The mean number of misses made by all the subjects in the second experiment when looking at the spheres from the end view was 14.97. Their mean number of misses from the side view was 8.48. The difference between these means was 6.49. This difference is 5.31 times its standard error.

This difference between means would indicate that the discrimination of moving objects is a much easier task when these objects are viewed from the side than when they are viewed from the end or coming toward the subject.

The Pearsonian r was computed between the score made on the end view and the score made on the side view. This breakdown resulted in two groups. The first group consisted of subjects who took the order side view first, end view second. The second group consisted of those who took the order end view first, side view second. The resulting Pearson r 's were .07 and .31 respectively. In view of the size of these coefficients, it seems probable that different factors are involved when performing this task from the side and from the end.

Correlation with Vision Tests. In order to show possible relationship between the sphere test scores and visual skills as measured by

the Ortho-Rater, correlation coefficients were computed.

Both the Pearsonian r and an eta were computed for all the acuities and the far depth perception since it was not known in advance whether the correlation would be linear or curvilinear. In the case of the phoria scores both lateral and vertical, far and near, previous experience had shown that the relationship between success on a task and scores on these tests was curvilinear and therefore only etas were computed.

The Pearsonian r 's varied from $-.09$ to $.48$. The correlation ratios varied from $.24$ to $.79$. Since the correlation ratio is not a highly reliable measure where the number of cases is small the correction suggested by Guilford was applied. This correction resulted in coefficients varying from $.00$ to $.73$. The majority of the coefficients are positive. We must not forget, however, that the number of cases used in this experiment is relatively small and that any generalization based on a careful study of the correlation coefficients will not be highly valid.

It would appear, however, that the visual abilities of the subject as measured by the Ortho-Rater tests are correlated in a low but positive fashion with ability in discriminating moving objects.

Summary

A test of ability to discriminate fine details in moving objects was constructed. This test involved discriminating between marked and unmarked clear glass spheres as they were rolling down an incline. Two test conditions were investigated: 1. With the spheres rolling toward the subject; and 2. With the spheres rolling laterally past the subject.

This technique was found to have a reliability of .89 to .95 when the spheres were rolling toward the subject and .89 when the spheres were rolling past the subject.

It is felt that studies of this type might well reveal desirable changes in design of industrial machines to take account more effectively of the demands made upon the machine operator. Biomechanics has been concerned with the abilities and skills required in the operation of machines which are already installed and operating. It would appear that

relatively simple laboratory studies of the psychological and physiological implications while the machine is still in the design stage could make contributions resulting in equipment which, when finally installed, would be more efficient through an increase in the efficiency to be expected from the operator. The present study suggests the possibility of such an approach.

Scores on the moving object test were correlated with the scores in the visual skills tests measured by the Ortho-Rater. These correlations would indicate that there is a low but relatively consistent positive relationship between ability to differentiate fine details in moving objects and visual skill test scores.

Received April 25, 1949.

An Analysis of Visual Requirements in Industry *

E. J. McCormick

Occupational Research Center, Purdue University

The importance of certain visual skills to performance on industrial jobs has been emphasized by the results reported from various investigations such as those conducted by Tiffin (6, 7, 8, 9, 10), Kephart (2, 3), and Stump (4, 5). To a considerable extent these and other investigations have been made with respect to employees on individual jobs; they have revealed in specific situations marked relationships between visual skills and various aspects of performance such as production, turnover, and accident experience. The several studies made on individual jobs have suggested the possible desirability of investigating the relationship between certain visual skills and performance on a variety of industrial jobs.

Statement of the Problem

The Occupational Research Center of the Division of Education and Applied Psychology, Purdue University, is engaged in the development of vision-test profiles for various industrial jobs. These profiles are used by industrial and business organizations for the dual purpose of aiding in employee selection and of identifying those individuals whose visual skills may not meet the general visual demands of their jobs and who might therefore benefit from professional eye care.

In the past the visual profiles for jobs were individually adapted for specific jobs on the basis of the relationships for present employees between their visual skills as measured by the Ortho-Rater¹ and their criterion measurements. Some such relationships might be due to chance factors, however, since the vagaries of human

traits are of such a character that in almost any single situation it is likely, by chance alone, that *some* statistically significant relationship might show up; if such a relationship *appears* to be a logical one, there is the temptation to accept it without cross-validation on a "hold-out" group.

The present investigation is an extension of research on the relationship between certain visual skills and job performance, directed toward providing a basis for establishing visual-skill profiles for industrial use which would avoid the potential pit-falls of individually adapted profiles. In the light of this general objective two specific objectives were established, namely: (1) The examination of the relationship between visual acuity and job performance of employees on a number of different jobs in order to determine any basic relationships that are of general applicability; and (2) The possible development of a different method for establishing visual acuity test cut-offs for visual profiles for certain types of industrial jobs.

Basic Data

The basic data used in the study were the results of Ortho-Rater vision tests and measures of job performance on approximately 5500 employees on 92 jobs in a number of industrial establishments. This information was obtained from the extensive files of the Occupational Research Center.

Selection of Hold-Out Jobs. A sample of 51 jobs that were in general randomly selected was drawn from files of the Occupational Research Center to serve as a "hold-out" group. This sample of "hold-out" jobs was drawn in order to be able to cross-validate on it the results of certain analyses (to be described later) that were made on a group of "base" jobs. Only jobs with about 20 or more employees were included.

The 51 jobs were then empirically divided into two groups on the basis of the judged character of the predominant visual demands. The first group (Group 1) included predominantly near vision jobs—those which were considered to require close, and rather constant, visual attention within a relatively restricted range,

* This article is based on the author's dissertation of the same title submitted to the Faculty of Purdue University in partial fulfillment of the requirements for the degree of Doctor of Philosophy, August, 1948. The dissertation was directed by Dr. N. C. Kephart.

¹ The Ortho-Rater is a visual testing instrument manufactured by the Bausch and Lomb Optical Company, Rochester, N. Y., for the visual classification and placement of industrial employees. A description of these tests is available elsewhere (11).

usually within arm's reach. Typical of the jobs included in this group are hosiery loopers, sewing-machine operators, assemblers of small parts, and certain types of visual inspectors.

The other group (Group 2) included jobs which required varying degrees of both near and far visual attention, and might therefore be considered as combination near and far vision jobs. Such jobs typically require some relatively close visual attention, say within arm's reach, as well as some farther visual attention at varying distances, either within or beyond the immediate work environment. Illustrative of jobs included in this group are knitting machine operators, weavers, spinners, and punch press operators.

Selection of Base Jobs. In order to make certain types of analyses, a sample of similarly representative "base" jobs was selected. A careful empirical examination was made of the relationship between visual acuity and the performance criteria of employees on these jobs. The only jobs retained were those on which it was empirically ascertained that there was a satisfactory degree of relationship between visual acuity and job performance. (These base jobs were so "selected" in order to ascertain, by subsequent analyses, the patterns of acuity skills which presumably contributed to performance on such jobs, with the thought that such "patterns" could later be cross-validated on the hold-out jobs.) A total of 41 such base jobs was retained. These base jobs likewise were divided into Group 1 (near vision jobs) and Group 2 (combination near and far vision jobs).

Criteria of Job Performance. The criteria of job performance varied from plant to plant and from job to job, and were based on such factors as production data or earnings, or were the results of ratings, including some ratings made by the paired comparison technique. In most instances the criterion measurements had been reduced to four or five numerical categories. The criterion categories for each job were divided into two groups, usually of approximately the same size, which were then identified as "high criterion" employees and "low criterion" employees. In instances where there was a middle category which precluded an approximately even division of the individuals into two groups, the middle criterion category was omitted.

Vision Tests. Ortho-Rater vision-test scores were available on all employees of the "base" and "hold-out" jobs. The Ortho-Rater includes tests of the following visual skills: (1) Far tests (given at the optical equivalent of 26 feet): phoria, vertical; phoria, lateral; acuity, both eyes; acuity, right eye; acuity, left eye; depth perception; and color discrimination; and (2) Near tests (given at the optical equivalent of 13 inches): acuity, both eyes; acuity, right eye; acuity, left eye; phoria, vertical; and phoria, lateral.

The six acuity tests were given particular attention in the present investigation. The results of the far acuity, both-eye test and of the near acuity, both-eye test were used directly. The results of the far acuity tests of the right and left eyes were compared for each individual and only the lower, or "worse-eye" test score, was used. A similar comparison was made of the near acuity tests for the right and left eye. The

four measures of visual acuity actually used were then: far acuity, both eyes; far acuity, worse eye; near acuity, both eyes; and near acuity, worse eye.

The range of possible scores on these tests is from zero to 15.²

Relationship Between Visual Acuity and Job Performance

The over-all relationship between each of the four visual acuity tests and job performance is reflected by the proportion of individuals scoring at each score level who were high criterion employees. This basic relationship for each test is presented graphically in Figures 1, 2, 3, and 4, which include the relationship for all of the 51 hold-out jobs combined, as well as for the Group 1 jobs and Group 2 jobs separately.

Both the far and near acuity, both-eye graphs show (Figures 1 and 3), for all 51 jobs combined, a relatively marked increase in the proportion of high criterion employees at each successive score value in the lower score range, with a less marked, though still persistent, increase at each successive score level in the higher score range. When considering the Group 1 and Group 2 jobs separately, however, it will be observed for the far acuity, both-eye test, that the Group 1 jobs do not reflect the drop at the lower end of the test-score range that occurs for the Group 2 jobs. For the near acuity, both-eye test, both Group 1 and Group 2 have essentially the same pattern of relationship with job performance.

For the two worse-eye tests (Figures 2 and 4), the positive, relatively straight-line, relationship from the low to high scores indicates that with higher worse-eye acuity (both near and far) the probabilities of satisfactory performance on jobs of the types studied increases quite constantly. A slight up-swing for the Group 1 jobs at the very low scores, especially for near acuity, worse eye, implies that for constant, close visual work a *marked* restriction of acuity is more conducive to job success than a *relatively low* level of acuity, presumably because the worse eye does not then hamper the effective use of the better eye as it might if the acuity level were only relatively low.

² For a conversion of these scores to other common measures of visual acuity the reader is referred to the manual of standard practice (11).

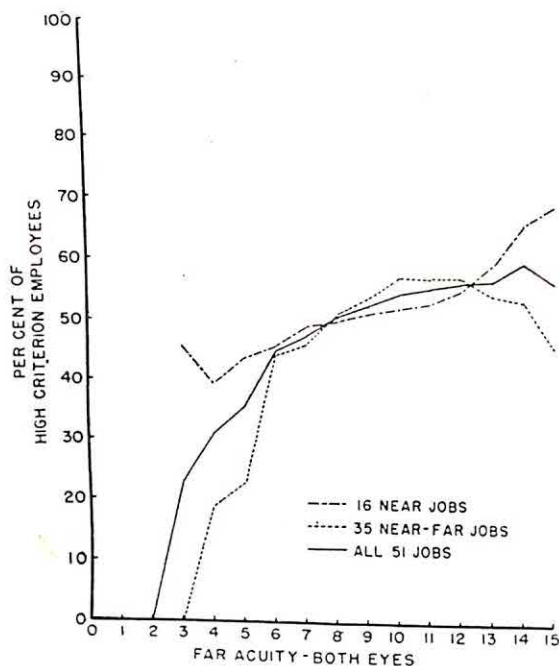


Fig. 1. Per cent of high criterion employees to total employees at successive scores on Far Acuity, Both-Eye Test.

Aside from this difference between Group 1 and Group 2, however, the curves for the two groups are basically the same.

Experimental Vision Test Profiles

Previous mention has been made of some of the potential disadvantages of developing visual standards for individual jobs on the basis of the relationship between the vision test scores and the performance criteria of employees on the jobs in question. To try to overcome these disadvantages, two different methods were developed for setting visual-acuity standards for jobs of the types investigated. These methods resulted in the establishment of acuity-test cut-off scores which were incorporated in "profiles" for the jobs.

Acuity Profile Derived from "Average" Acuity Scores. For each of the 41 base jobs a visual-acuity profile was individually developed which resulted in an adequate degree of differentiation between the high criterion and low criterion employees. For each of the four acuity tests, the average of the Ortho-Rater cut-off scores which were incorporated in these individually-developed profiles was determined.

The whole scores nearest the averages for the respective tests were established as the cut-off scores for the first type of experimental profile. This profile (designated as profile A) included the following cut-off scores: far acuity, both eyes, 8; far acuity, worse eye, 6; near acuity, both eyes, 8; and near acuity, worse eye, 7.

Acuity Profiles Adjusted to the Visual Acuity Level of Employees on the Job. The second type of experimental acuity profile was designed to establish the cut-off scores on the four acuity tests on the basis of some standardized relationship to the general acuity level of the employees on the job in question. This "standardized relationship" was developed by an analysis of data on the base jobs, for subsequent application and cross-validation on the hold-out jobs. The specific procedures involved in the analysis of the 41 base jobs are given below.

1. Procedures used in analyzing data on each of four acuity tests (Note: these are described in terms of a single test, near acuity, both eyes, but were similar for all four tests):
 - a. For each of the 41 base jobs, a frequency distribution was made of the scores of the employees on the near acuity, both-eye test.

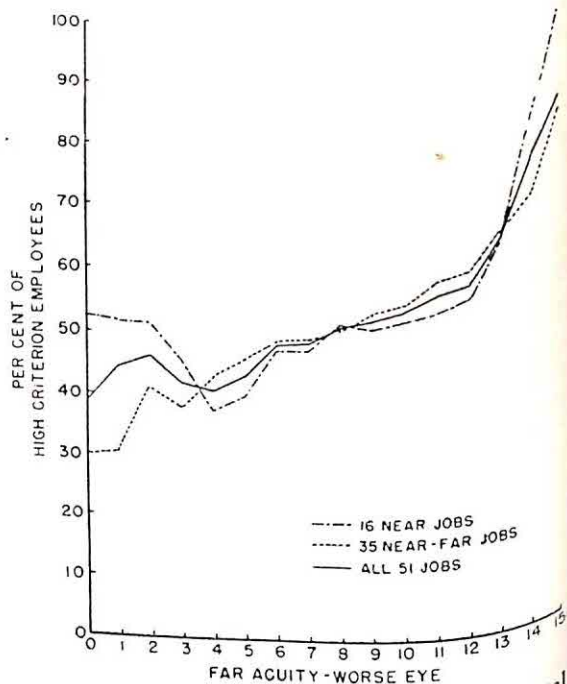


Fig. 2. Per cent of high criterion employees to total employees at successive scores on Far Acuity, Worse-Eye Test.

b. For each of the 41 frequency distributions certain percentiles were computed to the nearest whole-scores. The percentiles used were the 10th, 15th, 20th, 25th, and 30th, since a preliminary investigation suggested that the percentiles in this range would most adequately serve the purposes of subsequent analyses.

c. For each job an individualized visual acuity profile had been previously developed (with cut-off scores on the four acuity tests) which was empirically judged to differentiate adequately between the high criterion and the low criterion employees.

d. For each job the difference in score units was determined between the both-eye near acuity cut-off score of the previously developed profile and each of the specified percentiles on the test of the employees on the job.

e. These differences in score units were then summarized for all 41 jobs, there being a separate summary of these differences for the 10th, 15th, 20th, 25th, and 30th percentiles. Independent summaries of the same type were also made for the Group 1 jobs and for the Group 2 jobs. Varying degrees of "stability"

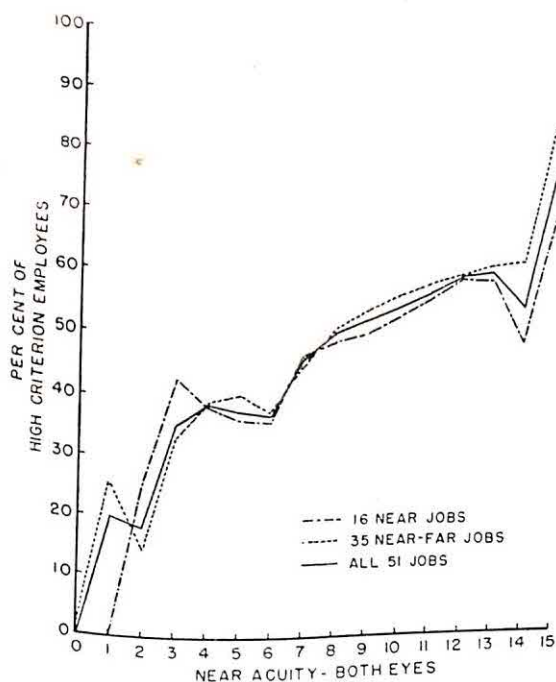


Fig. 3. Per cent of high criterion employees to total employees at successive scores on Near Acuity, Both-Eye Test.

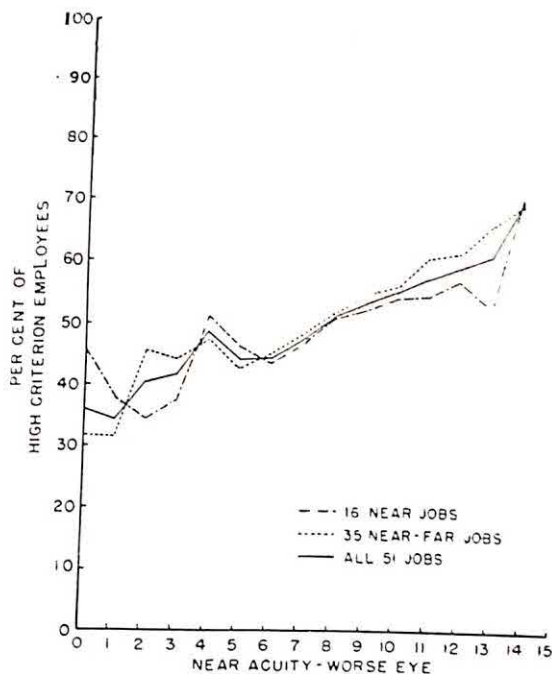


Fig. 4. Per cent of high criterion employees to total employees at successive scores on Near Acuity, Worse-Eye Test.

of these differences were noted. For the Group 1 jobs, for example, it was observed that for 15 of 20 jobs the cut-off score (of the individually developed profiles) was one score unit below the 25th percentile for the respective jobs. It was found that corresponding differences for some percentiles were less stable.

2. Procedures used in setting up formulas for experimental profiles of four acuity tests:

a. Several experimental profile "formulas" were developed on the basis of the type of analysis outlined above. These profile formulas embodied for the four acuity tests the more "stable" of the relationships reflected in the differences between the cut-off scores of the individually adapted profiles of the base jobs and certain percentiles. Following is an illustration (for one test) of the way in which these relationships were converted into profile formulas for use in establishing cut-off scores on the various tests for individual jobs: using the example given in (e) above for the near acuity, both-eye test, the cut-off score on this test for a given job would be set at one score unit below the 25th percentile of the distribution of scores of the employees on this test.

For the several experimental profiles of this type the cut-off scores on the four acuity tests for a given job would be established in a similar manner; various combinations of such relationships were used in the different experimental profiles. Some profiles were developed for use on jobs of the type included in Group 1, and others for jobs of the type included in Group 2. These profiles were designated by letter, from B through K.

b. Some of the experimental profiles of this type excluded certain tests, and some specified lower limits for cut-off scores (either in the form of specified percentiles or specific scores). These experimental variations were suggested on the basis of more detailed analysis of the data on the 41 base jobs.

Profiles of Phoria, Depth, and Color Tests. In order to examine the comparative results of the various experimental acuity profiles by themselves, and also when phoria, depth, and color tests were added to the battery, an auxiliary profile which included a specified combination of cut-off scores on these tests was added to all of the acuity profiles. This auxiliary profile (profile X) was developed in conjunction with a companion study (I), and consisted of the test cut-off scores which were most frequently included in the profiles of the base jobs as originally established in the regular procedures used by the Occupational Research Center.³ The combination of the X profile with an acuity profile is indicated by the addition of X to the alphabetical identification of the acuity profile, as BX, or KX.

Application of Experimental Profiles

The various experimental profiles, developed from data on the *base jobs*, were applied to the *hold-out* job groups as indicated below for the purpose of cross-validation: Group 1 jobs: profiles A, AX, B, BX, C, CX, D, DX, E, EX, F, FX, G, GX, H, HX; and Group

³ Profile X included the following cut-off scores:

Test	Score
Far vertical phoria, and near vertical phoria	
Low-end cut-off (left hyperphoria)	1
High-end cut-off (right hyperphoria)	9
Far lateral phoria, and near lateral phoria	
Low-end cut-off (esophoria)	2
High-end cut-off (exophoria)	14
Depth Perception	1
Color Discrimination	1

2 jobs: profiles A, AX, B, BX, I, IX, J, JX, K, KX.

The "A" and "AX" profiles included fixed cut-off scores on the various tests. For the other profiles (B through K and BX through KX) the acuity-test cut-off scores for the individual jobs were set by the formulas previously described.

For each profile applied to a given job the test scores of each employee were examined, and each individual who failed one or more of the tests (whose test score on a given test was at or below the cut-off score for that test) was considered as failing the profile.

For each profile applied to each job, the numbers of high criterion employees and of low criterion employees who passed and who failed the profile were determined. The proportions of those passing and of those failing each profile were then determined, and the critical ratio of the difference between these proportions was computed.

The subsequent step involved the determination of the effectiveness of each profile on the entire group of jobs to which it was applied. For this purpose an over-all critical ratio was computed.⁴

Results of Experimental Profiles

Table 1 presents the results of the several experimental profiles on the entire group of jobs to which they were individually applied. This table includes for each profile the over-all critical ratio previously mentioned. This over-all critical ratio for a given profile implies the probability that the differentiation between the high criterion and low criterion employees could be due to chance factors.

Acuity Profile. When examining the over-all critical ratios of the acuity profiles for the hold-out jobs it will be observed that

⁴ The over-all critical ratio for each profile was computed from the following formula applicable for replications of similar experiments, as provided by Dr. I. W. Burr, Department of Mathematics, Purdue University:

$$CR_a = \frac{M_{CR_i}}{\sigma_{M_{CR_i}}} = \frac{M_{CR_i}}{\frac{1}{\sqrt{N}}}$$

in which CR_a stands for the critical ratio for all jobs combined, M_{CR_i} stands for the mean of the critical ratios of the individual jobs, and N stands for the number of jobs.

Table 1

Summary of Results of Experimental Vision Test Profiles on 51 Hold-out Jobs

Profile	Over-all CR	Profile	Over-all CR
<i>Group 1 (16 near vision jobs)</i>			
A	2.46	AX	4.13
B	2.03	BX	4.23
C	2.35	CX	5.04
D	1.69	DX	4.08
E	2.86	EX	4.26
F	2.81	FX	4.08
G	1.97	GX	4.17
H	3.52	HX	5.16
<i>Group 2 (35 combination near and far vision jobs)</i>			
A	7.07	AX	5.95
B	8.75	BX	6.21
I	7.31	IX	5.59
J	9.14	JX	6.31
K	8.10	KX	6.08
<i>Group 1 and 2 (51 Jobs)</i>			
A	7.24	AX	7.24
B	8.39	BX	7.51

they range from 1.69 (5 per cent confidence level) to 9.14, which is far beyond chance expectations, with most of them being above the 2 per cent level (2.33). In terms of the over-all critical ratios, therefore, most of the acuity profiles may be considered as producing a differentiation that is beyond reasonable chance expectations.

The various acuity profiles applied to the Group 1 jobs all produced essentially the same results on that group; similarly, the several profiles applied to the Group 2 jobs all gave approximately the same results on this group of jobs. In general, however, the results for the Group 2 jobs are more pronounced than for the Group 1 jobs. It was because of this difference in over-all results that more experimental profiles were developed and applied to the Group 1 jobs than to the Group 2 jobs.

With regard to the results of the profiles on the Group 1 jobs, however, profile H (which excludes the far acuity tests and provides more rigid standards for the near acuity tests) resulted in a somewhat greater degree of differentiation than the other profiles. This implication of the relatively restricted importance of far acuity on such jobs (in comparison with

near acuity) is in accord with the over-all relationships previously discussed.

The degree of differentiation between high criterion and low criterion employees that resulted from the experimental profiles might logically be considered as reflecting a minimum degree of the true relationship between the visual skills and job performance. Among the several possible factors that might contribute to this under-evaluation are the character of the criteria for some jobs, possible differences in illumination, differences in specific duties of various individuals included in the same job category, differences in age, and the possibility that certain jobs reflect a restricted range of talent due to the possible elimination after placement of some of the visually unfit.

It is also pointed out that in the case of five of the Group 1 jobs and five of the Group 2 jobs it was ascertained empirically that not even an individually adapted job profile would result in a reasonable degree of differentiation, perhaps due to inadequate criteria or chance factors. For such jobs it therefore could not be expected that a standardized profile or profile formula would result in a statistically significant differentiation.

Profiles Including Acuity, Phoria, Depth, and Color Tests. In addition to the application of the various acuity profiles by themselves, the acuity profiles were also individually applied with the addition of the auxiliary X profile, which includes a set of fixed scores for the phoria, depth, and color tests. The addition of the X profile to the acuity profiles applied to the Group 1 hold-out jobs consistently resulted in additional discrimination.

With the Group 2 jobs, however, the addition of the X profile to the various acuity profiles resulted in a moderate decrease from the results obtained from the acuity profiles by themselves, although even then the over-all critical ratios were from 5.59 to 6.31. While this actual observed decrease was traced primarily to the systematic influence of a negative relationship for a very few specific jobs, there does not appear to be any evidence that the addition of the phoria, depth, and color tests adds particularly to the differentiation on Group 2 jobs over that obtainable with the acuity tests by themselves. Presumably the acuity tests tapped a good share of the meas-

urable relationships between visual skills and performance on the jobs in this group.

Since the phoria, depth, and color tests apparently contribute to the differentiation of Group 1 jobs over that produced by the acuity profiles, one inference that might logically be made is that in the case of jobs which require relatively constant and close visual attention, the adequate adjustment of convergence as measured by the phoria tests, and possibly the ability to perceive depth relationships, presumably are related to satisfactory performance on such jobs.

Implications

The results of the investigation suggest certain factors which might be considered in the establishment of visual standards for jobs of the type studied.

For example, the basic relationships between the four visual acuity tests and job performance reflected in Figures 1, 2, 3, and 4 imply that there should be a moderate lower limit for near acuity in both eyes for both near vision jobs (Group 1) and for combination near and far vision jobs (Group 2), and also a moderate lower limit for far acuity for combination near and far vision jobs (Group 2). Aside from these moderate minima, however, it appears that the visual requirements for jobs of the types studied are relative rather than absolute; it would seem then that various specific visual standards which require relatively adequate degrees of acuity may be expected to produce, in the over-all, relatively satisfactory results in differentiating between high criterion and low criterion employees on jobs of the types investigated.

In addition, it would seem that for most effective results the visual acuity standards for a given job should in some way be adjusted to the visual acuity level of employees on the job; this implication is derived from the fact that the experimental profiles that were so adjusted tended to produce a somewhat greater degree of differentiation between the high criterion and low criterion employees than did the fixed-score acuity profile (profile A) which provided the same set of cut-off scores for all jobs.

With regard to the possible use of standardized profile formulas for setting visual acuity

job-standards, it should be pointed out that while no such formula can be expected to produce the degree of differentiation *among present employees* that is possible by developing individually adapted profiles for specific jobs, nevertheless, the use of standardized profile formulas seems to offer a more adequate basis, *in the long run*, for predicting job performance on most jobs of the types investigated and for the referral of employees for professional eye care.

Summary and Conclusions

An investigation was made of the relationship between certain visual skills and performance on various types of industrial jobs, using Ortho-Rater scores of approximately 5500 employees on 41 "base" jobs and 51 "hold-out" jobs, and measures of job performance which identified high criterion employees and low criterion employees. The investigation was made with particular attention to four measures of visual acuity, namely, far and near acuity in both eyes, and far and near acuity in the worse eye. The investigation included two phases: first, an analysis for each of the four tests of the proportion of all individuals on the hold-out jobs who scored at each score level who were high criterion employees; and second, the development and application of several experimental vision test profiles and an analysis of their results.

The primary conclusions are as follows:

1. Visual acuity requirements for satisfactory performance on jobs of the types studied seem to be general and relative rather than specific and absolute.
2. A moderate minimum of near visual acuity in both eyes is especially pertinent to satisfactory performance for both "near" and "combination near-far" jobs, and a moderate minimum of far visual acuity is equally pertinent to performance on "combination near-far" jobs; beyond such minima, performance increases quite constantly with greater degrees of acuity, but at a more gradual rate.
3. The increase in the probability of job success with increases in worse-eye acuity (both near and far) is relatively constant over the entire range of scores.
4. The several experimental acuity profiles (developed from data derived from the base

jobs and applied to the hold-out jobs) gave substantially similar results on the group of jobs to which they were individually applied (one group of near vision jobs and another group of combination near and far vision jobs).

5. The results of practically all the profiles were beyond reasonable chance expectations, although the results were more pronounced on the combination near and far jobs than on near jobs.

6. The addition to each of the acuity profiles of a standard set of fixed scores on phoria, depth, and color tests contributed noticeably to the differentiation between high criterion and low criterion employees on the near jobs, but not on the combination near and far jobs.

7. The possible use of standardized profiles shows definite promise of providing, in the long run, relatively satisfactory standards of visual skills for jobs of the types investigated.

Received May 6, 1949.

References

1. Carr, E. R. *An analysis of the relationship of phoria, depth perception and color discrimination to job performance*. Master's thesis, Purdue Univ., 1948.
2. Kephart, N. C. Visual skills and labor turnover. *J. appl. Psychol.*, 1948, **32**, 51-55.
3. Kephart, N. C. An analysis of professional eye care and industrial efficiency. *Trans. Amer. Acad. Ophthal. & Otol.*, March-April, 1946, 166-178.
4. Stump, N. F. Spotting accident-prone workers by vision tests. *Fact. Mgmt. & Maint.*, June, 1945.
5. Stump, N. F. Visual functions as related to accident-proneness. *Personnel*, 1944, **21**, 50-56.
6. Tiffin, J. *Industrial psychology*. (2nd Ed.) New York: Prentice-Hall, Inc., 1947.
7. Tiffin, J. Vision and industrial production. *Illum. Engng.*, 1945, **40**, 230-257.
8. Tiffin, J. The use of visual data as an aid to increase production and efficiency. *Trans. Amer. Acad. Ophthal. & Otol.*, January-February, 1944.
9. Tiffin, J., and Greenly, R. J. Employee selection tests for electrical fixture assemblers and radio assemblers. *J. appl. Psychol.*, 1939, **23**, 240-263.
10. Tiffin, J., and Rogers, H. B. The selection and training of inspectors. *Personnel*, 1941, **18**, 14-31.
11. Standard practice in the administration of the Bausch and Lomb occupational vision tests with the Ortho-Rater. Bausch and Lomb Optical Company, Rochester, N. Y., 1944.

The Effect of Ordinal Position upon Responses to Items in a Check List

Donald T. Campbell and Phillip J. Mohr

Departments of Psychology and Speech, The Ohio State University

With the increased use of questionnaires and check lists has come a healthy awareness of the many possible sources of bias in such instruments. One source of bias, which this paper considers, is position effect. Two types of position effect may be noted: (1) the effect of ordinal position, or position *per se*; and (2) the effect of specific sequences, involving content interaction. The effect of ordinal position has been mentioned in standard works on the questionnaire method (e.g. 2) and this possibility has been regarded by some professional survey workers to be of sufficient importance to necessitate systematic controls.

This anxiety is justified by surprisingly little published research. In factual tests, where responses presumably are based on knowledge, studies have shown significant ordinal differences for the alternatives of a multiple choice item, although with somewhat inconsistent results (1, 9, 11). Mathews has similarly studied the effect of different left-right arrangements of response alternatives for the items of an interest test (12). In opinion poll studies fragmentary reports indicate possible position effects (4; 5, pp. 34-5; 7). Of these, Cantril (5) finds significant but inconsistent differences for various orderings of two or three alternatives to a single poll question. From the available abstracts of the two other studies, it would appear that specific sequence effects (in which content interaction might be involved) are confounded with purely ordinal effects (such as might be due to primacy, recency, fatigue, or the like). No studies have been located in which the effect of position *per se* has been studied systematically.

Our study attempts to isolate the effect of ordinal position in a preference check list already in use in professional work. We borrow from the annual Iowa Radio Audience Survey. The director, Dr. F. L. Whan, has been one of those who have been worried by the possibility of position effect, and has attempted

to control it by using in rotation multiple forms of a radio program types list. With Dr. Whan's permission, we have taken with slight modification the check list of 16 radio program types used in the 1946 survey. Modifications in procedure have been introduced only insofar as necessitated by group administration in college classes.

The check list, as we used it, had the following instructions:

"Listed below are sixteen general types of radio program materials. In the spaces provided at the right, check the *FIVE* types that you like *BEST*. Please check *ONLY FIVE*—no more, no less."

Half of the questionnaires listed the 16 types with examples provided for each type. The examples, as in Dr. Whan's studies, represented well-known popular programs with comparatively high "Hooper" ratings. The other half of the questionnaire gave no examples. We wanted to determine whether or not relative ambiguity, i.e., absence of examples, would accentuate any position effect, on the assumption that if the stimuli (the program types) were less well defined respondents might be more apt to choose with reference to position of an item on the check list rather than to its content. The 16 program types are listed below, with the asterisk indicating the termination of the items in the "without examples" series:

1. Livestock and Grain Market Reports*
2. Talks on Farming and Farm Problems*
3. News Broadcasts*, including local news, network commentators and farm news
4. Talks and Discussions of Important Public Affairs*, such as talks by congressmen, American Town Meeting of the Air, etc.
5. Devotional Programs*: Sermons, talks on religion, etc.
6. Hymns, Religious Music*, such as Hymns of All Churches, Salt Lake City Tabernacle Choir, etc.
7. Classical or Semi-Classical Music*, such as Met-

	Farm Markets	Farm Talks	News	Public Affairs	Devotional	Hymns	Classical Music	Western Music	Popular Music	Brass Bands	Comedians	Variety Programs	Quiz Shows	Dramatic Shows	Serials	Sports Events
1.	I	M	G	F	K	D	A	H	B	C	E	N	L	O	J	P
2.	B	I	C	M	E	G	N	F	L	K	O	D	J	A	P	H
3.	D	N	I	A	L	M	H	O	G	B	J	F	C	P	K	E
4.	L	B	K	I	O	C	D	M	J	E	A	G	P	N	H	F
5.	A	O	F	E	I	H	J	K	N	M	B	P	D	L	G	C
6.	G	D	B	N	J	I	F	A	C	L	P	M	K	H	E	O
7.	K	C	J	G	H	L	I	D	E	P	F	B	O	M	A	N
8.	J	L	E	B	A	K	G	I	P	O	N	C	H	D	F	M
9.	M	F	D	H	C	N	O	P	I	G	K	A	B	E	L	J
10.	N	A	M	O	P	F	P	E	D	I	L	H	G	J	C	K
11.	O	E	H	K	M	P	L	C	A	F	I	J	N	B	D	G
12.	C	G	L	D	P	B	M	N	K	J	H	I	E	F	O	A
13.	F	H	N	P	G	A	E	J	M	D	C	O	I	K	B	L
14.	E	K	P	C	F	J	B	G	O	H	M	L	A	I	N	D
15.	H	P	A	J	D	O	K	L	F	N	G	E	M	C	I	B
16.	P	J	O	L	N	E	C	B	H	A	D	K	F	G	M	I

FIG. 1. The latin square design (letters designate questionnaire forms).

- ropolitan Opera, New York Philharmonic Orchestra, Andre Kostelanetz, etc.
8. Old-Time Pioneer or Western Music*, such as National Barn Dance, etc.
 9. Popular Music and Popular Orchestras*, such as Fred Waring, Lucky Strike Hit Parade, Vaughn Monroe, etc.
 10. Brass Bands*, such as the U. S. Army, Navy, or Marine Bands, etc.
 11. Comedians*, such as Jack Benny, Bob Hope, Fred Allen, etc.
 12. Variety Programs*, without featured comedians, such as Breakfast Club, Arthur Godfrey's Talent Scouts, etc.
 13. Quiz or Audience Participation Programs*, such as Break the Bank, Quiz Kids, Vox Pop, Truth or Consequences, etc.
 14. Complete Dramatic Shows*, such as Lux Radio Theatre, Screen Guild, etc.
 15. Daily Continued Story Serials*, such as Ma Perkins, When a Girl Marries, Lum and Abner, etc.
 16. Broadcasts of Sports Events*, such as broadcasts of football, basketball, baseball games, fights, etc.

Experimental Design. A "latin square" provides the ideal design for such a study (8), but to our knowledge it has not previously been used for this purpose. Through this device, we prepared sixteen check list forms, with each item appearing once on each form, and once in each ordinal position. Subject to these restrictions, the distribution of items in

the latin square might well have been random. But Fisher (6, pp. 267-269) points out that "in a well-planned experiment certain restrictions may be imposed on the random arrangement of the plots in such a way that the experimental error may still be accurately estimated." As mentioned before, this study was designed to measure the effect of ordinal position,—apart from sequence of content interaction effect. Therefore, rather than using a completely random latin square, we used one in which the sequence was systematically controlled by having each program type preceded once and followed once by every other type.¹

The latin square is shown in Figure 1. The letters A through P identify the sixteen basic questionnaire forms. For example, in form A, Classical Music was first, Dramatic Shows second, Public Affairs third, etc. Because of the use of the "with examples" series and the "without examples" series, a total of 32 different forms was actually used.

¹ The writers are indebted to Dr. D. R. Whitney and the Statistics Laboratory of the Department of Mathematics, The Ohio State University, for providing this "homogeneous" latin square and for doing the bulk of the calculations. In passing, it may be noted that squares of this sort are available only for even numbers (3). The Statistics Laboratory will be happy to provide such squares to those interested.

The Sample. The questionnaire was submitted to 1280 students in 35 beginning classes in Psychology and Speech. Forty students filled in each form. Eight hundred and eighty-four of the respondents were men, 374 women, with 22 neglecting to indicate sex.

Administration of the Survey. No advance warning was given to the students. Each instructor passed out the questionnaires, with these instructions. "I am sure this is self-explanatory. Do not collaborate with your neighbors. Do not ask about or mention aloud any program or specific radio program, since that might influence the response of others." We are reasonably sure that the students felt that this was simply a survey of radio program preferences of college students.

The 32 forms were distributed randomly in all classes, each respondent's ballot being different in arrangement of choices from those of his neighbors.

Results

The Analysis of Results. The results are presented in Figure 2, combining cases with and without examples. They are presented graphically in Figure 3, keeping these two

sets of data separate. In Figure 2, the frequency of the choice of each program type is given for each of the 16 positions. For example, 40 respondents selected "Broadcasts of Sports Events" (No. 16) when that item appeared in the 1st position; 37, when it appeared in the 2nd position, etc. The maximum possible for any program type in any given position is 80. The totals at the bottom indicate the frequency of choice of the program types; the totals at the right, the frequency of choice of the ordinal positions. Note that the frequency of choice of each program type in each position is summed in the marginal totals at the bottom; in the totals at the right, the frequency of choice of each position for each program type is summed. In short, each position and each program type had equal opportunity to be chosen. Note also that all 1280 respondents had the opportunity of being represented in each marginal total—for position and for program types—thus equating samples for all comparisons.

As might be expected, significant differences occur among the preferences for the 16 program types. "Popular Music and Popular Or-

	Farm Markets	Farm Talks	News	Public Affairs	Devotional	Hymns	Classical Music	Western Music	Popular Music	Brass Bands	Comedians	Variety Programs	Quiz Shows	Dramatic Shows	Serials	Sports Events
1.	2	44	19	3	5	51	1	73	18	59	32	31	48	1	40	427
2.	5	37	31	3	6	50	6	68	10	62	30	38	42	2	37	429
3.	1	60	19	1	1	49	4	70	10	63	22	26	47	2	45	422
4.	1	2	51	15	3	—	40	1	69	6	58	23	19	44	1	37
5.	—	5	45	22	2	4	37	—	72	9	54	29	33	42	4	48
6.	1	2	38	21	3	4	43	1	70	7	59	30	33	46	—	38
7.	3	2	40	21	3	3	47	4	72	8	60	24	31	49	2	38
8.	1	3	43	20	3	6	48	3	68	7	55	16	28	43	3	42
9.	1	2	46	18	1	5	36	6	73	9	56	20	23	42	2	48
10.	4	4	43	27	—	4	50	4	70	12	63	25	30	42	1	45
11.	—	2	51	23	—	2	38	2	71	8	54	24	31	52	1	39
12.	2	3	43	19	2	6	45	3	65	3	56	20	24	42	2	43
13.	—	1	35	21	2	5	48	4	63	6	65	22	23	44	1	44
14.	1	1	43	24	—	5	55	3	70	7	52	25	27	45	2	45
15.	1	3	42	21	3	4	45	4	72	5	51	24	30	42	—	39
16.	1	1	41	25	2	5	46	1	72	12	51	13	37	46	1	37
TOTAL	20	39	702	346	31	65	728	47	1118	137	918	379	464	716	25	665

FIG. 2. Choices obtained (with and without examples combined).

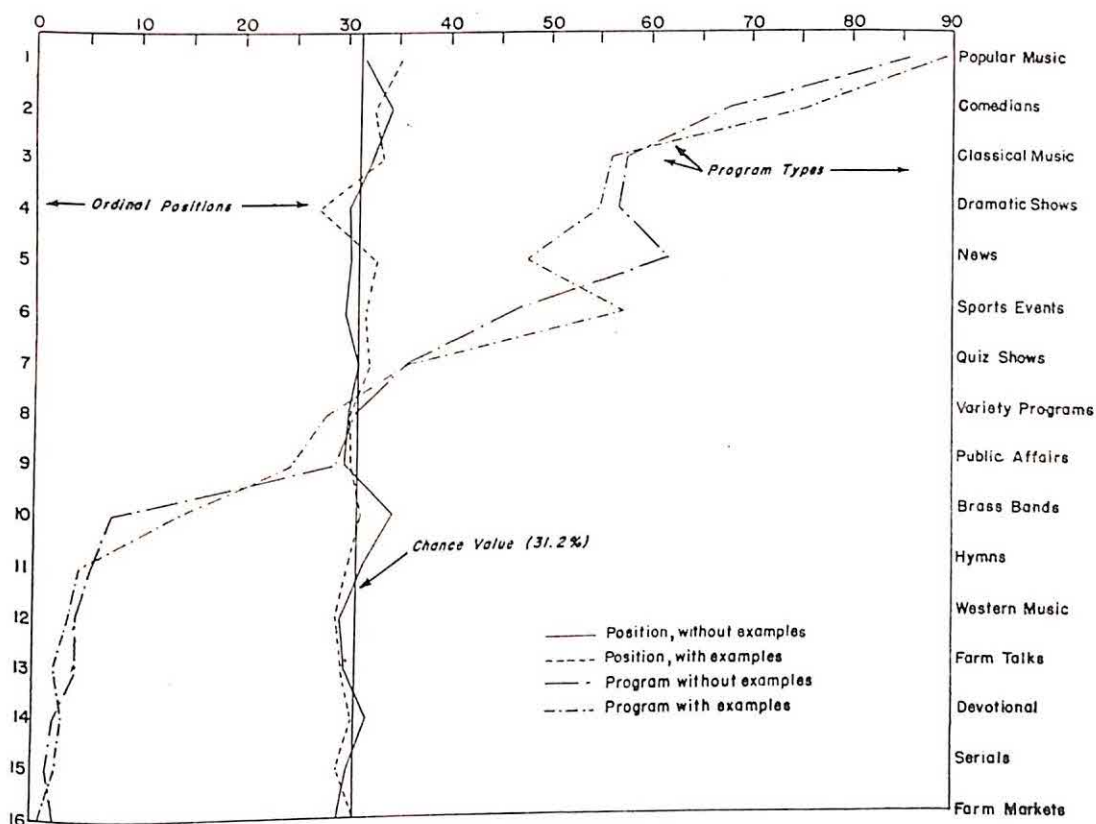


FIG. 3. Percentage of choice by position and by program type.

chestras" was most preferred; "Livestock and Grain Market Reports" (Farm Markets) was least preferred. But to the surprise of the experimenters no apparent position effect is found.

The graphic presentation of the material in Figure 3 supports this. This chart superimposes the data for program preference and the data for position effect, to present visually the contrast in variability provided by the two dimensions of the data. In Figure 3 there is a separate presentation of data for "with examples" and "without examples," which data were presented combined in Figure 2. Note the lack of any greater position effect in the "without examples" series. Both lines seem to vary at random about the expected value of 31.2% (400/1280). The presence of examples does seem to have had an effect on the popularity of certain programs.

Statistical tests of significance confirm in general the absence of position effect. The marginal totals for the ordinal positions in

Figure 2 were tested collectively by Chi Square against the expected value of 400 which would obtain if position had no effect. The obtained Chi Square of 12.4 had a P-value of .68, indicating that values as large or larger than 12.4 could occur *by chance* 68 times out of 100. The hypothesis that differences among the totals are but chance deviations from the expected value cannot be rejected by this test. Similar tests were made of the position effect for each program type separately (each column in Figure 2) and none of these Chi Squares reached a 5% level of significance. If *any* position effect is present, it is not sufficiently strong to manifest itself with this number of cases, and as tested by Chi Square.²

² Our use of a Fisher-type design in collecting the data may lead to a question as to why analysis of variance statistics have not been applied. In passing it may be noted that latin squares and simultaneous use of several classification criteria do not lend themselves exclusively to analysis of variance. The present data, being enumeration data, do not meet the basic assumptions of analysis of variance. While an arc-sine transformation of the data might have partially overcome this weak-

Our original expectations of a position effect involved, of course, some notion of *trend*. More specifically, we had expected the first positions to receive more choices than other positions, with a downward trend from first to last positions. *Visual inspection* of Figures 2 and 3 offers little support for this notion. Note that, for the combined data, position 2 is highest and position 4 is lowest; position 10 one of the highest, and position 12 one of the lowest. However, a more complete answer to our initial notion of trend was obtained from an application of Mann's "T" test. This non-parametric statistic of trend seemed especially applicable for our purposes in that it measures upward or downward trends in data that are enumerative in character, without the requirement of equality of intervals between steps, and making no assumptions of normalcy. The application of this test to our data indicates that, at a very marginal level of significance ($P = .048$), the totals at the right in Figure 2 tend to decrease with an increase in ordinal position number. This level of significance certainly is not very high, particularly when the relatively large number of cases is considered. And, as noted above, frequent exceptions occur to this trend, if it is a trend.

A further source of evidence on the presence or absence of a position effect is found in the comparison of the position totals from different samples. If position were to have a highly complicated and non-linear effect, it should none the less be consistent from sample to sample. With this in mind, the correlation coefficients among position totals for four subsamples were computed, as shown below. "(A)" represents the series of questionnaires which provided examples for each program type; "(B)", the series which gave no examples:

ness, in the present data there is still another limitation. Because of the nature of the assignment to the respondents (to pick 5 of 16 alternatives), totals on certain classification criteria are fixed at equality in advance (e.g., comparisons involving with and without examples, sex, and questionnaire forms). The Statistics Laboratory did exploratory analyses of variance on total and sub-populations, with program type and position as classification criteria. In no instance did *position* reach the 5% level of significance.

Product Moment Correlation of:

Men.....(A) with Men.....(B) is	.21
Women....(A) with Women....(B) is	-.05
Men.....(A) with Women....(B) is	.14
Women....(A) with Men.....(B) is	.26
Men.....(A) with Women....(A) is	-.05
Men.....(B) with Women....(B) is	.13

These correlations describe little, if any, similarity in the obtained position totals from one sample to another. Note, however, that the general tendency is for the values to be positive, with two exceptions. Once again, there is no clear evidence for a position effect, but only some slight indications that one may be present.

The present data seem to indicate little if any effect of ordinal position upon the number of choices received by the items in the check list. It is to be emphasized that the conditions under which these results have been obtained are highly unique and specific. We would suggest no generalizations of these results, obtained on college students, to the legendary housewife whose soup is boiling over on the stove during the interview. Similarly, our results may be specific to the longish 16 item check list, or to the strict assignment of five choices, or the content of the items. Further research is obviously needed. Each commercial user of the check list may want to determine the presence or absence of a position effect under his own operating conditions. We do believe, however, that through the use of designs such as the one presented in this study, the definitive answer for the specific situation can be obtained with the relatively small effort of preparing systematically multiple forms of the check list.

Received April 18, 1949.

References

1. Atwell, E. R., and Wells, F. L. Wide range multiple choice vocabulary test. *J. appl. Psychol.*, 1937, 21, 550-555.
2. Blankenship, A. B. *Consumer and opinion research*. New York: Harper and Brothers, 1943.
3. Bugelski, B. R. A note on Grant's discussion of the latin square principle in the design and analysis of psychological experiments. *Psychol. Bull.*, 1949, 46, 49-50.

4. Cahalan, D., and Tamulonis, V. M. The effect of question variations in public opinion surveys. *Amer. Psychol.*, 1947, 2, 328 (Abstract).
5. Cantuil, Hadley. *Gauging public opinion*. Princeton: Princeton University Press, 1944, 34-35.
6. Fisher, R. A. *Statistical methods for research workers*. New York: Hafner Publishing Co., 1948.
7. Flowerman, S. Polls on anti-semitism: an experiment in validity. *Amer. Psychol.*, 1947, 2, 328 (Abstract).
8. Grant, D. A. The latin square principle in the design and analysis of psychological experiments. *Psychol. Bull.*, 1948, 45, 427, 442.
9. McNamara, W. J., and Weitzman, E. The effect of choice placement on the difficulty of multiple-choice questions. *J. educ. Psychol.*, 1945, 36, 103-113.
10. Mann, Henry B. Non-parametric tests against trend. *Econometrica*, 1945, 13, 245-259.
11. Mathews, C. O. The effect of position of children's answers to questions in two-response types of tests. *J. educ. Psychol.*, 1927, 18, 445-457.
12. Mathews, C. O. The effect of the order of printed response words on an interest questionnaire. *J. educ. Psychol.*, 1929, 20, 128-134.

Identification of Cola Beverages. IV. Postscript

N. H. Pronko and D. T. Herman

University of Wichita

Three earlier studies¹ on the identification of cola beverages have led to the present investigation. The first study employed Coca Cola, Pepsi Cola, RC Cola, and Vess Cola. Administered to 108 Ss, results showed an almost equal distribution of identification responses among the first three categories. Since S's refused to use the fourth name, it was decided to use only the first three drinks in a second study on the hypothesis that identification responses would be more nearly equally distributed among the three beverage names. The hypothesis was substantiated as was also a subsequent one in Study No. 3 which employed three unknown or less well-known colas but which showed an almost chance distribution of the Coca Cola, Pepsi Cola and RC Cola names to the three "dark horses." On the assumption that subjects might do better if they were told to identify Coca Cola, Pepsi Cola and RC Cola when these beverages were actually administered and when the subjects were told that these were the Colas they were to identify, the present study was undertaken.

Procedure

As in the previous studies, two groups of subjects were used—105 Ss in Part I and 60 in Part II. These were beginning students in Elementary Psychology courses.

Part I. Each of 105 Ss was admitted individually into the experimental room and was asked to sit down, after which the following instructions were read to him:

"We would like to have you taste and identify some cola drinks. We will have three colas: Pepsi Cola, Coca Cola and RC Cola. You will be told in what order and when you are to drink them. After you have finished each sample report your identification to me. After each stimulus presentation, take enough water from the paper cup before you to rinse your mouth well."

¹ Pronko, N. H., and Bowles, J. W., Jr. Identification of Cola beverages. I. First study. *J. appl. Psychol.*, 1948, 32, 304-312; II. A further study. *J. appl. Psychol.*, 1948, 32, 559-564; III. A final study. *J. appl. Psychol.*, 1949, 33, 605-608.

A tray containing three one-oz. glasses of Coca Cola, Pepsi Cola and RC Cola respectively was placed before the S. He was then instructed in what order to drink the beverages labelled X, Y and Z. Samplings were spaced about a minute apart during which interval pertinent data were recorded.

The order in which the three beverages were presented was determined pre-experimentally and was such that each beverage was administered in the first, second and third positions an equal number of times (i.e., 35 times). This counterbalanced order was intended to nullify position effects or taste interactions orally. All beverages were kept out of sight of Ss and were refrigerated at approximately 5° C.

Part II. In Part II, 60 Ss were administered the same Cola drink at each of three trials. Thus, 20 got all Coca Cola; 20, all Pepsi Cola; and 20, RC Cola. In all other respects, including the instructions, the procedure was the same as that for Part I.

Results

Results show that of 105 Coca Cola samples, 57 are correctly identified but 48 are misidentified. In the case of the 105 Pepsi Cola samples 45 are correctly labelled and 60 misidentified. This is similar to the case of RC Cola which is correctly identified 47 times and misidentified 58 times. In all three cases, these colas are misidentified almost as often as they are correctly identified.

In Part II (where each of 20 Ss was given three samples of the same cola) there was a frequency variation of from 17 to 23 correct responses.

As to the percentage of correct responses when Ss were given different colas, for the Coca Cola samples, there were 54% correct identifications, 43% for Pepsi Cola and 45% for RC Cola. As in the previous study, Coca Cola is in the lead as regards correct identifications. For all three brands 34% are correctly called and 66% incorrectly.

A comparison of critical ratio tests of the hypothesis that the various identification responses are not on the basis of actual taste stimuli was then made. The correct identifications of the three respective Colas do not show the same results. Those for Pepsi Cola and RC Cola are low, actually 1.44 and 1.71 respectively. However, the CR for the Coca Cola category is 3.19 which indicates a statistically significant difference between the obtained frequency and chance expectancy. Furthermore, analysis of data based on frequency of correct responses shows very low ratios ranging from 0 to 1.00, indicating that in no instance does any obtained frequency vary significantly from chance expectancy.

Next the percentages of correct identifications of Parts I and II were compared. In the former case three different Colas were given while in the latter three identical samples were administered. In the case of Pepsi Cola and RC Cola, the tests of a statistically significant difference between Parts I and II yield the low critical ratios of .76 and .89 respectively. However, the CR obtained for a difference between Coca Cola identified correctly in Parts I and II is 3.44, a statistically significant difference which suggests a behavioral difference between these parts. Apparently, as a group, our subjects do not discriminate in the same way in the two situations. There is a difference in the number of correct discriminations when they are given three samples of the same Cola as compared with administration of three different Colas.

Significance tests of a comparison of Study II where correct names were not given in the instructions to S and the present study in which they were indicated, show that when the frequency of correct identifications for the two studies is compared, no reliable differences are found. However, this does not mean that Coca Cola is not being identified with better than chance frequency as found in the present study.

A similar comparison for Parts II of Study II and the present study yields CR's for differences in Coca Cola, Pepsi Cola and RC Cola

correct identifications that are not reliably different. Those for the incorrect identifications show high variability.

The overall results from this study suggest that when Ss are placed in the comparatively restricted situation of the present experiment, their identifications are no better than in Part II of the previous studies (i.e. when they are administered three identical samples of a cola beverage). Their identification responses are not statistically different from chance expectancy. The same holds for identification of the Pepsi Cola and RC Cola drinks. The only shift concerns the Coca Cola beverage which is identified with greater than chance frequency in this situation where choice of correct name is experimentally limited to the three names given the subject in the instructions. Narrowing his choice apparently permits him to make more strikes, although even in this situation he misidentifies Coca Cola almost as often as he identifies it.

Summary

A total group of 165 Ss was asked to identify one-oz. samples of the following three Cola Beverages: Coca Cola, Pepsi Cola and Royal Crown (RC) Cola. In Part I, 105 Ss were presented one of each of three different Colas while in Part II, 60 Ss were given three samples of the *same* beverage being evenly divided among the three different classes. Instructions to both groups stated the beverages to be the three named above.

In general, when Ss are given three samples of the *same* beverage their identifications are not significantly different than when the beverages are unknown Colas or actually these Colas but unspecified. The same holds for Pepsi Cola and RC Cola identifications when three different samples are given. The situation with Coca Cola is different, for in this study this beverage is identified with a frequency that yields a statistically significant difference from chance expectancy.

Received October 21, 1949.
Early publication.

Book Reviews

Reynolds, Lloyd G., and Shister, Joseph. *Job horizons*. New York: Harper and Brothers, 1949. Pp. x+102. \$2.25.

This is a descriptive report to lay readers issued by the Labor and Management Center of Yale University. It is concerned with causes of labor mobility or immobility as given to fifteen interviewers by 800 manual workers in a New England manufacturing city in 1947.

The first two chapters cover the general problem and importance of mobility and the factors in job satisfaction. The factors are defined and discussed in an interesting manner and include numerous quotations from manual workers. The next two chapters relate to how workers go about finding new jobs, what causes them to pick one job rather than another, why they left school, and how they went about getting their first job. Few persons will fail to profit from a careful reading of these chapters, and they are of special value to vocational counselors. Emphasis is placed on the non-rational procedure typically followed by applicants, their lack of knowledge of the labor market, and inappropriateness of current economic theory. The last two chapters cover movement up the occupational ladder (which is obviously limited because the study is restricted to those in manual occupations) and the worker's view of job opportunity. This last chapter draws the other chapters together and points out that the worker's behavior is nonrational or irrational from the viewpoint of economic theory rather than the circumstances as the worker sees them.

The book is concluded with a brief appendix on methods including sampling. Many readers will question the use of samples randomly selected from manual workers listed in city directories. As the authors have pointed out, many of the more mobile workers have been eliminated.

As a research report, even for the general public, this book is deficient in several ways. Unwarranted generalizations and conclusions seem to have been made, and the book has no index or bibliography. Many of the findings

which appear revelatory to the economist-authors will be considered commonplace by psychologists. No reference is made to other work in the field of worker motivation, job satisfaction, or job preferences. Many of these other researches, apparently conducted with care equal to this one, conflict with some of the findings reported here. Neither such conflicts (nor the agreements) are mentioned. This is unfortunate for the lay as well as technically trained reader. The book frequently refers to "workers." It is to be regretted that the more restricted "manual workers" was not invariably used to prevent the reader from over generalization.

Nevertheless, this small book is fascinating reading, and is packed with interesting facts and views. Few readers will finish the book without mental stimulation and new ideas.

C. E. Jurgensen

Minneapolis Gas Company.

Tobias Wagner. *Selective job placement*. New York: National Conservation Bureau, Association of Casualty and Surety Executives, 1946. Pp. 151.

The author of this little volume has compiled a number of tables and charts from the comparative study of the work efficiency of a group of normal workers and a group of workers with various types of disabilities. These comparisons are presented in the form of averages, e.g., average number of days absent per year, average rate of production, average rated quality of production, etc. Much of the potential value of the book as a source of information on the industrial usefulness of the disabled is lost because the author has failed to present two very important types of statistical information. At no point in the book does the author state the number of cases of either normal or disabled workers concerned in this research. Despite the bristling appearance of the book in terms of the statistical data presented, no indications of the reliability of the averages nor any test of the significance differences of the averages is ever presented. These

are serious shortcomings in any research work, and the omission is difficult to understand.

In general, the approach to the topic of worker placement, whether disabled or non-disabled, is rather elementary. If Dr. Wagner was writing this book for the lay person, his looseness of description and generality of statement might be forgiven, but he speaks of directing this book toward "personnel managers, safety engineers, rehabilitation specialists, and employers in general."

The general impression given by the author is that he is selling the merits of disabled workers to the industrial employer. However, it does appear that the disabled are at least as efficient in industrial jobs where their disability is not a handicap, as the normals in the same jobs.

The book might well be used in industrial concerns where attempts are being made to get management to accept disabled people for employment. It could be used as an argument for the disabled. It does not, however, seem to have any place in the college curriculum as a text, since the contribution of the book is rather limited.

The somewhat repetitious style of the author makes continued reading a little difficult, as the same statements or ideas are presented at the beginning and the end of consecutive chapters. The frequent use of such terms as always, every, entirely, absolutely, impossible, constantly, absolutely necessary, all, and completely, are often used ill-advisedly.

The book has merit in bringing together in one volume the research conducted by Dr. Wagner, and despite its brevity, it covers a great deal of ground. The section on the types of physical disabilities reads like a medical dictionary, and covers the orthopedic, visual, and hearing disabilities. Despite the inadequacy of the reported statistics, this book would serve as a useful reference on this specific topic.

A. A. Canfield

University of Southern California

Flesch, Rudolf *The art of readable writing*.
New York: Harper and Brothers, 1949.
Pp. 237. \$3.00.

"To come right out with it," writes the author, "this is a book on rhetoric. Its pur-

pose is to help you in writing." With this introduction Dr. Flesch attempts to present a "modern, scientific rhetoric" for informal, useful, every-day writing. While this reviewer would be presumptive to evaluate the rhetoric *per se*, it is his opinion that this book is an excellent contribution to more effective communication.

The Art of Readable Writing takes a wider view of the writing process than its predecessor, *The Art of Plain Talk*. While the first book was primarily concerned with the readability of sentences and words, in the second book the whole technique of writing readably is surveyed.

In chronological sequence Flesch takes up the problems which a writer faces (or should face) in writing. In clear short chapters, packed with well-chosen examples to illustrate his meanings, he follows the writing effort from the initial idea to the final draft. How-to-start is dealt with by concrete suggestions for audience appraisal, collecting facts, assimilating data and getting a "slant" or theme. What-to-say is covered by recommendations for meaningful introductions, narrative style and personalizing the message. How-to-say-it is given with instructions on writing colloquially, shaking off some misconceptions about form and content, and achieving the correct level of difficulty and interest.

In addition Flesch warns his readers about misunderstandings arising from ambiguous words and reader errors. He includes an especially timely caution concerning the use of diagrams and pictures to convey information without adequate written explanation.

An appendix explains the use of the revised readability yardsticks essentially as given in *J. appl. Psychol.*, 1948, 32, 221-233. Nomographs for computing reading ease and human interest scores are printed on the end papers of the book. A good, annotated reading list for more specific help in writing and a complete set of notes on the illustrative and scientific references for each chapter round out the book.

The presentation is skillful, readable and entertaining throughout. While the work may not be as "scientific" as Flesch implies it is, it is certainly an excellent combination of current knowledge with the shrewd insight and "know-how" of a readability expert who can really write.

This book is one which should be read and kept on the desk of everyone who writes to be understood. It has implications for all areas of writing: letters, news stories, textbooks, handbooks, advertisements or union contracts. It is a book of high interest and real importance to communicating people whether they be in academic or applied fields.

James J. Jenkins

University of Minnesota

Williamson, E. G. (Editor). *Trends in student personnel work*. Minneapolis: The University of Minnesota Press, 1949. Pp. x+417. \$5.00.

Too often the papers presented at a conference commemorating some individual or institution sound like the minutes of a mutual admiration society when later published. This volume of forty-three papers by forty authors, although originally given in November, 1947 at the University of Minnesota to celebrate 25 years of personnel work there, is a notable exception for several reasons: (1) as the title indicates, the authors were as much concerned with the future steps to be taken as with the accomplishments of the past; (2) the breadth of the Minnesota program as reviewed insured a comprehensive treatment of the total field; and (3) the long period of leadership in the field made the event of general as well as local interest.

Following an appropriate introduction, the papers deal successively with the development of student personnel work in all its major aspects. Part I. The Role of Personnel Work sets the tone of the conference by showing, in articles by Willey, Cowley, MacLean, and McClusky, how the personnel point-of-view arose and its necessity for constructing a sound educational program for American youth. Following a survey (Part II) of aptitude and interest and personality testing by Stuit and Darley, respectively, in Part III, Paterson demonstrates vocational counseling utilizing the best of available case-study and measurement techniques and Shartle discusses areas of research necessary to understand occupational adjustment.

Part IV, by Bordin and Porter, helps the reader to understand the *directive versus non-*

directive controversy in the light of the basic issues and long-term trends involved. Subsequent parts deal with the more specialized areas of student personnel work including mental hygiene, problems of veterans, women students, and foreign students, marriage counseling, discipline, speech, financial problems, medical services, housing. The problems of selection, training and utilizing faculty counselors are given extensive treatment as is religious counseling, as viewed by representatives of the major faiths. In an integrated series of papers, Lloyd-Jones, Wrenn, and Darley trace the history of personnel work as a profession and present a realistic appraisal of its present stage of development. Although emphasized throughout, the importance of the social and cultural influences of the campus upon the student's total growth is pointed up in papers by Cowley and Sutherland. In the concluding Part XIV, Turner and Tyler discuss recent developments in achievement testing and their effect upon college admission policies and curriculum construction. References at the end of each paper and a detail index of the volume increase its value as a source-book.

This publication should have a wide readership. Administrators will find it a powerful stimulus to the development of student personnel programs and faculty will be given an understanding of the rationale and functions of student personnel work. Personnel workers will find many suggestions for improving their own programs and personal effectiveness and students training for personnel work will find it a useful reference book and source of ideas for significant research. Although not designed as a textbook in the ordinary sense, it can readily serve as the basic source-book for a graduate seminar in student personnel administration.

Albert S. Thompson

Teachers College,
Columbia University

Brouwer, Paul J. *Student personnel services in general education*. Washington, D. C.: American Council on Education, 1949. Pp. 317. \$3.50.

Student Personnel Services in General Education, covers the period from 1939 to 1944. It includes a description of participation by an

original group of 22 colleges and universities minus 7 drop-outs, plus three late additions. It is a bewildering book. Basically, it is a report of the findings about *student personnel services* offered by an extremely heterogeneous group of colleges sampled to include "the land-grant college, the municipal university, the state teachers college, the independent liberal arts college, the Catholic college, the Protestant church-related college, the Negro college, the four-year college for women, the junior college for women, and coeducational junior college." Some of its best parts are those which reflect the thinking of a particular author, rather than the agreements of the group which participated in the program.

The heterogeneity of participating institutions, plus individual viewpoints in the writing of various sections, contribute to unevenness in writing, and some lack of coherence. Unfortunately, there is neither a subject-matter nor author index. These omissions offer the usual handicap which drives the interested but baffled reader to frenzy.

The reviewer found two sections of particular interest. These are Section 1, *The counseling service* (pp. 7-46) and Section 5, *Pre- and post-college personnel services* (pp. 102-114). The first has been written quite recently, because it contains some excellent writing covering the extremes of the continuum of counseling methodology. Some of the materials would not have been available if they had been written at the time the program was in progress. The author almost makes a distinction between *counseling* and *advising*. Had this distinction been clearly made, it would have done much to clarify confusion in this regard in the minds of many college administrators and college instructors. The term *counseling* is encountered as a process carried on by students. Those readers who struggled through graduate courses and degrees to reach the first

rung on the professional counseling ladder must be excused for the slight shudder.

Section 5, *Pre- and post-college personnel services*, is weak in contrast to Section 1. A real question is whether the legitimate objective of friendliness and rapport is not sufficient to be gained from these methods without making the questionable claim that we know the student sufficiently well through casual interview and unverified information to help him materially in his planning.

Part III, *The principles of personnel services* (Sister Annette) is, in the opinion of the reviewer, the heart of the book for those who are interested in counseling as the foundation of student personnel work. Sister Annette's recognition of differences in the basic individual personality structures of teachers and its effect on what they can do as personnel workers is refreshing (p. 242).

General criticisms of the book include failure to give adequate attention to biological heredity, to special aptitudes and abilities, to mental organization as a basis for curriculum building, to trait variability, and to multiple curricula. Despite these shortcomings, parts of the book are valuable to an extent that causes the reviewer to recommend it for purchase. It is hoped that the policies of the American Council on Education will encourage more specificity in its publications treating student personnel work than has been the case in this publication. If some of the topics in this book had been treated in terms of institutional enrollment differentials and privately *versus* publicly supported colleges, the publication might have been far more valuable to a larger group of readers. And, finally, can't we have a law which requires an index for all books?

Milton E. Hahn

University of California, Los Angeles

New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor,
Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota

- Pavlov, a biography.* B. P. Babkin. Chicago: University of Chicago Press, 1949. Pp. 470. \$6.00.
- Interaction process analysis.* Robert F. Bales. Cambridge: Addison-Wesley Press, Inc., 1949. Pp. 224. \$6.00.
- Workbook in personnel methods.* Robert M. Bellows and Carl H. Rush, Jr. Dubuque: Wm. C. Brown Co., 1949. Pp. 102. \$2.10.
- Conference methods in industry.* Henry M. Busch. New York: Harper and Brothers, 1949. Pp. 107.
- Applied experimental psychology.* Alphonse Chapanis, Wendell R. Garner, and Clifford T. Morgan. New York: John Wiley and Sons, Inc., 1949. Pp. 434. \$4.50.
- In the name of common sense.* Revised edition. Matthew N. Chappell. New York: The Macmillan Co., 1949. Pp. 172. \$2.75.
- Essentials of psychological testing.* Lee J. Cronbach. New York: Harper and Brothers, 1949. Pp. 475. \$4.50.
- Current trends in industrial psychology.* Wayne Dennis et al. Pittsburgh: University of Pittsburgh Press, 1949. Pp. 198. \$3.75.
- Courts on trial.* Jerome Frank. Princeton: Princeton University Press, 1949. Pp. 441. \$5.00.
- Psychoanalysis.* Second edition. Edward Glover. New York: Staples Press, Inc., 1949. Pp. 367. \$4.00.
- The prediction of categories from measurements: with applications to personnel selection and clinical prognosis.* J. P. Guilford and William B. Michael. Beverly Hills: Sheridan Supply Co., 1949. Pp. 55. \$1.40.
- Organization of behavior.* D. O. Hebb. New York: John Wiley and Sons, Inc., 1949. Pp. 335. \$4.00.
- Trends in industrial relations.* Bulletin No. 16. Alexander R. Heron et al. Pasadena: Industrial Relations Section, California Institute of Technology, 1949. Pp. 88. \$1.00.
- Bristow Rogers: American Negro.* Else P. Hillpern, Irving A. Spaulding, and Edmund P. Hillpern. New York: Hermitage House, Inc., 1949. Pp. 184. \$3.00.
- Leadership and isolation.* Helen Hall Jennings. New York: Longmans, Green, and Co., Inc., 1949. Pp. 240. \$3.00.
- Therapeutic group work with children.* Gisela Konopka. Minneapolis: University of Minnesota Press, 1949. Pp. 134. \$2.50.
- The magic cloak.* James Clark Moloney. Wakefield, Mass.: The Montrose Press, 1949. Pp. 345. \$5.00.
- Introduction to psychopathology.* Lawrence I. O'Kelly. New York: Prentice-Hall, Inc., 1949. Pp. 736. \$4.50.
- Industrial and occupational trends in national employment. 1910-1940, 1910-1948.* Research Report No. 11. Gladys L. Palmer and Ann Ratner. Philadelphia: Industrial Research Department, University of Pennsylvania, 1949. Pp. 80. \$1.00.
- Mass communications.* Wilbur Schramm, Editor. Urbana: University of Illinois Press, 1949. Pp. 552. \$4.50.
- Varieties of delinquent youth.* William H. Sheldon. New York: Harper and Brothers, 1949. Pp. 899. \$8.00.
- Predicting success in professional schools.* Dewey B. Stuit, Gwendolen S. Dickson, Thomas P. Jordan, and Lester Schloerb. Washington, D. C.: American Council on Education, 1949. Pp. 187. \$3.00.
- Foundations of method for secondary schools.* I. N. Thut and J. Raymond Gerberich. New York: McGraw-Hill Book Co., Inc., 1949. Pp. 493. \$4.00.
- Personality maladjustments and mental hygiene.* Second Edition. J. E. Wallace Wallin. New York: McGraw-Hill Book Co., Inc., 1949. Pp. 581. \$5.00.
- Hypnotherapy of war neuroses.* John G. Watkins. New York: The Ronald Press Co., 1949. Pp. 384. \$5.00.
- Sight, light and efficiency.* H. C. Weston. London: H. K. Lewis and Co., Ltd., 1949. Pp. 318. 42s. net.

Theory of hearing. Ernest Glen Wever. New York: John Wiley and Sons, Inc., 1949. Pp. 484. \$6.00.

Children's voluntary reading as an expression of individuality. Mary Hayden Bowen Wollner. New York: Bureau of Publications, Teachers College, Columbia University, 1949. Pp. 117. \$2.35.

Identifying and developing potential leaders. Personnel Series No. 127. New York: American Management Association, 1949. Pp. 39. \$75.

The Harvard list of books in psychology. Cambridge: Harvard University Press, 1949. Pp. 77. \$1.00.

Subscription Lists of the American Psychological Association

MEMBERS AND AFFILIATES

Approximately 9,300 names

The American Psychological Association maintains an address list of its members and affiliates, which is for sale providing the nature of its use is in conformity with the purposes of the Association.

1950 Prices

Envelopes addressed \$35.00

(advertiser furnishes envelopes and pays express charges)

Addresses on tape, not gummed \$17.50

(suitable for a mailing machine)

STATE LISTS

Priced according to number of names wanted

SUPPLEMENTARY LISTS

Approximately 3,500 names in total list

Individual journal lists vary from 400 to 1,600

The Association also maintains a list of subscribers who are not members of the Association (universities, libraries, industrial laboratories, hospitals, other types of institutions, and individual subscribers). The general list for all journals includes all types. Each single journal has a more specialized circulation.

For any one journal, envelopes addressed \$15.00

For any one journal, addresses on tape \$10.00

For all journals, envelopes addressed \$20.00

For all journals, addresses on tape \$15.00

For further information, write to

American Psychological Association

1515 Massachusetts Avenue Northwest

Washington 5, D. C.

Journal of Applied Psychology

VOL. 34, No. 2

APRIL, 1950

A Difficult New Test of Mechanical Comprehension *

William A. Owens, Jr.

Iowa State College

The Bennett Test of Mechanical Comprehension is undoubtedly the prototype among schematic, pictorial tests which place a premium upon mechanical "common sense" and insight into basic, physical principles. In a recent survey of V.A. Guidance Centers throughout the country it was reported in use in 90% of them, and in frequent use in 64% (1).

Since it is so generally employed, it is only natural that it should occasionally be utilized with populations differing rather markedly from that for which it was originally intended. Under these circumstances maximally satisfactory results can hardly be expected.

The writer has for some time been associated with a collegiate institution at which academic mortality in the Engineering Division is heavy, and in which there has been correspondingly great need for improved tools and techniques of selection. In early 1946 it became apparent that a revised and extended Bennett-type test, more schematic and more difficult than Form BB, might be a real asset in this context. Such a test has been constructed and will be briefly and incidentally described in what follows.

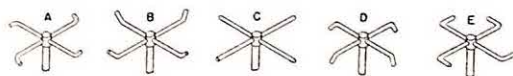
Problem

The central problem of the present investigation is to evaluate this new Test of Mechanical Comprehension, Form CC, in the potential selection of engineering students.

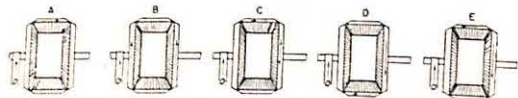
Materials

Figure 1 contains typical content from the new test. It may be observed that there are five response possibilities, instead of the three previously employed, and that the probability

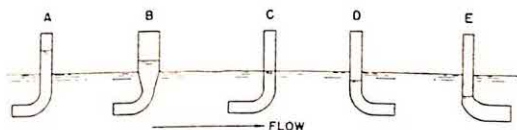
* The writer is deeply indebted to Dr. James Wert for statistical assistance and advice.



60. Which sprinkler will spin the fastest?
(1) A, (2) B, (3) C, (4) D, (5) E.



61. With the drive D as indicated, which picture correctly shows how the gears would turn?
(1) A, (2) B, (3) C, (4) D, (5) E.



62. Which picture shows a physical impossibility?
(1) A, (2) B, (3) C, (4) D, (5) E.

FIG. 1. Typical Content, Mechanical Comprehension Test, Form CC.

of chance success is thereby reduced. It may also be noted that the drawings are quite schematic, unembellished line drawings which assume a reasonable level of sophistication. Like Form BB, Form CC is an untimed power test.

Subjects

The subjects upon whose scores the succeeding results are based were all students in The Engineering Division at The Iowa State College. Freshman, Sophomore, and Senior classes are represented at various points, and the total number of cases tested is 725.

Methods

The data obtained were accorded the following types of treatments. 1. Test reliability,

within this relatively homogeneous group, was determined; and Mechanical Comprehension Test (M.C.T.) scores were correlated with scores on the A.C.E. Psychological Examination to estimate the relationship of the former to intelligence. 2. Hierarchies of scores made by groups presumably differing in the ability in question were tabulated. 3. Correlations of M.C.T. scores with certain selected course grades were determined. 4. A random sample of 321 freshmen tested in 1947 has now been followed longitudinally for a period of three years to determine the relationship of M.C.T. score to academic mortality. 5. Biserial correlations were computed between test score and "A" achievement in selected courses vs. the balance of the achievement distribution; and, similarly, between test score and "F" failure to achieve vs. the balance of the achievement distribution. An estimate of the efficiency of test prediction at the high and low ends of the achievement continuum has thus been obtained. 6. Multiple correlations were computed between the combined effects of A.C.E. score, high school average, and M.C.T. score and the criterion of selected course grades. Several estimates of the independent contribution of M.C.T. score have been obtained. 7. The importance of certain background factors, such as farm rearing and physics training, in the determination of M.C.T. score has been estimated, through application of analysis of variance and covariance techniques. 8. M.C.T. scores have been correlated with scores on the Kuder Preference Record in an

Table 1
M.C.T. Score and Educational Status

Group	Mean	Range	Inter Quartile Range	N
Agric. E. Freshmen	36	15-55	30-41	177
Engineering Freshmen	39	20-55	33-44	321
Archit. E. Sophomores	42	24-55	35-46	120
Engineering Seniors	47	32-57	42-51	107

attempt to uncover relevant factors of interest and motivation.

Results

The primary results of this study have been summarized in Tables 1 through 7. Prior to more detailed comments upon these it may be noted that the corrected odd-even reliability of the new M.C.T. in a population of 321 engineering freshmen was 0.80, and that scores on it correlated 0.39 with scores on the A.C.E.

Table 1, then, shows a hierarchy of M.C.T. scores which accords with expectation. It is apparent that educational attrition has eliminated many of the low scoring individuals. It is also suggested that the test is mainly an aptitude test in that, while medians rise through negative selection, only three of 107 seniors scored higher than the highest entering freshman—and that by mere 1 and 2 point margins.

Table 2 shows some representative correlations of the M.C.T. with course grades in various subject matter areas. Theoretical

Table 2
Correlations of M.C.T. Scores with Course Grades

	r	P	Engineering Group
1. Engineering Courses			
Theoretical and Applied Mechanics	0.49*	<.01	Eng. Seniors (N = 107)
Median grades in 7 relevant courses	0.41*	<.01	Eng. Seniors (N = 107)
2. Shop-Drawing			
Drawing and Projection	0.39	<.01	Arch. Eng. Sophs (N = 120)
Shop (estimated quality of product)	0.32	<.01	Ag. Eng. Sophs (N = 177)
Drafting	0.30	<.01	Eng. Freshmen (N = 260)
3. Math.-Science			
Chemistry	0.34	<.01	Eng. Freshmen (N = 260)
Algebra	0.28	<.01	Eng. Freshmen (N = 260)

* These coefficients are corrected for restriction in range of test scores.

Table 3

M.C.T. Scores and Mortality in Engineering

	N	Q ₃	Md	Q ₁
1. Passing	93	49	43	37
2. Failed	67	40	34	31
3. Dropped voluntarily	94	43	39	34
4. Changed divisions	57	45	40	33
5. Changed institutions	7	—	42	—
Total	= 318			

Passing vs. Failed, $r_{bis} = 0.59$; Failed vs. All Others, $r_{bis} = 0.46$.

and Applied Mechanics ($r=0.49$) involves the physical science subject matters of "statics" and "dynamics." Correlations with senior-level courses have been corrected for restriction in range of test scores because primary interest centers in the predictive value of the test as applied to freshman populations. Even so, the indicated relationships almost certainly provide underestimates of the corresponding true values, as the Registrar's Office at Iowa State College allows repetition of a course with only the *last* grade obtained appearing on the official record. While it is true that the correlation of test scores with shop grades is not high, two circumstances seem noteworthy: (1) There was only *one* experienced instructor in the group of five who estimated quality of shop product, and the test correlated 0.48 with his ratings; and (2) The A.C.E. and high school average correlate only 0.09 and 0.11, respectively, against this same shop criterion, thus making it appear that the M.C.T. predicts it relatively well.

Table 3 shows the relationship between the subsequent dispositions of 318 engineering

Table 4

M.C.T. Prediction by Achievement Level

Course	Entire Group	Low (F's) vs. Balance	High (A's) vs. Balance	N
Math. (Algebra)	0.28	0.30	0.41	260
Chem. (General)	0.34	0.26	0.48	260
English (Comp'n)	0.16	0.10	0.16	260

Combined probability that "Low" and "High" r's do not differ significantly = $<.01$.

freshmen tested during the Spring quarter of 1947, and the M.C.T. scores which they made at *that time*. All but three of the original group are accounted for; one has been killed and two have records too confused to be useful. The Table gives the status of all subjects as of mid-April, 1949, at which time those listed as "passing" were third quarter juniors. It may be noted that the pass-fail biserial correlation which appears below the table furnishes a most encouraging estimate of test validity, and that even if all "voluntary drops" and "transfers" be classed as "successes" the relationship with test score remains substantial.

Table 4 shows how achievement level relates to the efficiency with which the M.C.T. will predict grades in a sampling of freshman courses. As estimated from the 3 combined probabilities of a real difference, it appears that the correlations in Column 3 *are* significantly larger than those in Column 2, and that the test is, therefore, a better predictor of the grades of the *more* able students than it is of those of the less able. Since it was originally intended, in the construction of Form CC, to produce a test which would differentiate at *high* ability

Table 5

Prediction of M.C.T. with A.C.E. and High School Average (H.S.A.)

Course	(1) A.C.E.	(2) H.S.C.	(3) M.C.T.	A.C.E. H.S.A.	A.C.E. H.S.A. M.C.T.	Per Cent of Variance (1), (2) & (3)
Drawing (Engineering)	0.33	0.25	0.30	0.34*	0.39*	44-15-41
Chemistry (General)	0.50	0.53	0.34	0.58†	0.60†	33-47-20
1st Qr. Av.	0.55	0.47	0.37	0.58"	0.61"	47-30-23

† * " Probability of non-significant difference = $<.01$.

Table 6
Background Factors and M.C.T. Score

Characteristic	Highest Score	F	Chance P
Time on farm	Entire time	2.09	> .05
Father's occupation	Agriculture	7.60	< .01
Years of shop classes	More than one	1.25	> .05
Years shop experience	More than one	0.87	> .05
Years physics classes	One half year or more	11.47**	< .01
Age	25 years or more	0.05	> .05

** Score difference, Physics vs. Non-Physics means, is 4.5 points; initial ability controlled, via analysis of covariance, score difference is 2.5 points.

levels, these results are as desired and suggest that the stated purpose has been realized.

Table 5 shows that M.C.T. score makes a significant contribution to the prediction of first quarter grades, above and beyond those made in combination by score on the A.C.E. and by high school average. In each case the independent contribution of the M.C.T. is highly significant in spite of the fact that the criterion is not particularly appropriate. Column 6, at the extreme right, contains estimates of the percent of the total predictable variance in course grades contributed, successively, by the A.C.E., high school average, and the M.C.T.

The data for Table 6 were obtained by securing certain presumably relevant information about the backgrounds of 177 Agricultural Engineers who subsequently took the M.C.T. Responses to each item were assigned to one of two or three categories, and the test scores of the respondents were grouped accordingly; analysis of variance was then applied to estimate the significance of the intergroup differences in test scores. As indicated, only two of the six potential relationships were statistically significant. Confirming studies on urban-rural differences, subjects whose fathers were

engaged in some phase of agricultural work *did* have mean test scores about 3 raw score points higher than those whose fathers were otherwise classified occupationally. Likewise, subjects who had had one-half year or more of high school physics had mean test scores approximately 4.5 raw score points higher than those who had had less than this amount. It will, however, be immediately granted that there is some tendency for only the initially superior students to elect this difficult subject. Analysis of covariance was, therefore, applied to the results in the manner suggested by Lindquist (4) to control on inter-group differences in initial aptitude as evidenced by A.C.E. scores. This had the effect of reducing the difference between the adjusted physics and non-physics means to approximately 2.5 raw score points. Such a result is interesting in that mechanical comprehension tests have sometimes been criticized for their undue dependence upon formal physics training, apparently a relatively unwarranted criticism. Table 7 gives the intercorrelations of the M.C.T. and the Kuder Preference Record. It will be noted that the four starred coefficients differ significantly from zero, but insignificantly from each other. Some motivational influences may be indicated in the fact that subjects who are above average in scientific and mechanical interest tend to score higher on the M.C.T. It is also interesting to observe that the four significant relationships identified would fairly well define a typical engineering profile on the Kuder, except for the absence of a computational interest.

Table 7
Correlations of M.C.T. with Kuder P.R.
for 260 Students

Kuder Scales									
Me	Co	Sc	P	A	L	Mu	Ss	Cl	
.20*	-.03	.26*	-.07	-.02	-.05	.17*	-.11	-.19*	

* 1% $r = 0.16$.

Conclusions

The following conclusions seem warranted relative to the role of this new test of mechanical comprehension in engineering student selection.

1. In an engineering population, it yielded an odd-even reliability of 0.80 and correlated 0.39 with the A.C.E. Psychological Examination.
2. It yielded hierarchies of scores, by groups, which accorded with expectation.
3. It showed satisfactory correlation with relevant course grades.
4. It was an excellent predictor of academic mortality.
5. It predicted better at high achievement levels than at low.
6. In prediction of course grades, it made a significant independent contribution beyond that made by A.C.E. and high school average.
7. Score on it seemed to depend upon background of training or experience to the extent of only 2 or 3 raw score points.

8. It correlated with scores on the Kuder Preference Record about as would be expected of a test of mechanical or engineering aptitude.

Received July 8, 1949.

References

1. Baker, Gertrude, and Peatman, J. C. Tests used in Veterans Administration advisement units. *Amer. Psychologist*, 1947, 2, 99-102.
2. Bennett, G. K., and Cruickshank, R. M. *A summary of manual and mechanical ability tests*. New York: The Psychological Corporation, 1942. Pp. 75.
3. Kellum, W. E. Recent developments in selection of candidates for aviation training. *Amer. J. Psychiat.*, 1943, 100, 80-84.
4. Lindquist, E. F. *Statistical analysis in educational research*. New York: Houghton Mifflin Company, 1940. Pp. 257 and appendix.
5. McDaniel, J. W., and Reynolds, W. A. A study of the use of mechanical aptitude tests in the selection of trainees for mechanical occupations. *Educ. & Psychol. Measmt.*, 1944, 4, 191-197.
6. Traxler, A. E. Correlations between "mechanical aptitude" scores and "mechanical comprehension" scores. *Occupations*, 1943, 22, 42-43.

Study of Executive Leadership in Business.

III. Goal and Achievement Index *

C. G. Browne

Wayne University

The importance of the establishment of organizational goals, purposes, and objectives; the need for the effective communication of these established goals throughout the organization; and the evaluation of past achievements of the organization have been recognized by various writers on organization and executive activities. These factors are important because they are influential in decision-making, one of the characteristics of executive activity and leadership, which should be based on a purposive moving toward common objectives and which extends through all levels of supervisory activity. A method which can be used to study the effectiveness of these three functions in any given situation among any given group of employees will constitute a valuable addition to the total study of executive leadership and the eventual evaluation of executive performance.

Procedure

This article is devoted to the presentation of a method, hereinafter referred to as the Goal and Achievement Index, which has been applied on an exploratory basis to the study of 23 top executives in a tire and rubber manufacturing company. Previous articles have presented other methods used in the study (2, 3).

The subjects for the total study constituted all of the executives on the first, second, third, and fourth echelons of the business with the exception of one executive on the third echelon. They were classified into the following departmental groups: general administration, 4 cases; sales, 6;¹ finance, 4; manufacturing, 8; and personnel, 2.

The detailed procedure followed in devising the Goal and Achievement Index form and in

* The writer expresses his appreciation to Dr. C. L. Shartle for his original suggestion that a goal and achievement method be included in the explorations.

¹ By his choice, the Vice-President-Sales was not included in the Goal and Achievement Index data.

accumulating the data on it has been described elsewhere (1). However, the general procedure was as follows:

1. In a personal interview, each executive enumerated the goals of the company as he understood they existed in the planning of top management. The 96 goals enumerated then were classified into the five major departments of the business to which they best applied.

2. Seventeen representative statements were chosen and arranged in the same random order on two scales, one scale for the evaluation of the statements in terms of their importance as goals and the other for their evaluation in terms of the degree of company achievement in each.

3. The executives rated the statements on a five point rating scale which ranged from "extremely important" to "no importance" as goals and from "completely achieved" to "nothing done" in achievement terms.

4. In calculating the individual goal and achievement scores, a point value of from one to five was assigned to each of the rating columns, the lower point value indicating a higher rating column. The individual's score on each scale was the total points for the 17 statements as he rated them, a low score indicating higher goal importance and degree of achievement. To determine the rank order of the statements by departmental groups, the individual points for each statement of the executives in each departmental group were added, and the statements were arranged in rank order on the basis of the total number of departmental points for the statements.

Total Group Ratings

Table 1 includes the composite ratings of the total group of 23 cases, with the rank order of each statement for both goal importance and degree of achievement indicated, and the difference between the rank orders. Where the difference between goal and achievement was a plus value, it indicated that achievement was above its importance as a goal of the company, and when the difference was a minus value, the opinion was that the item had more importance as a goal than the company had thus far achieved with it.

The data in Table 1 can be analyzed in terms of departmental activities as indicated by the

Table 1

Rank Order of Composite Goal and Achievement Ratings for Total Group of 23 Executives

No.	Statement of Purpose	1* Dept.	2 Goal Rank Order	3 Ach. Rank Order	4 Diff. Between G & A Rank
1.	To create big production.	M	14	7	+ 7
2.	To make as much money as possible for the stockholders.	F	13	11	+ 2
3.	To continue to keep Congo a successful and prosperous company.	GA	1	6	- 5
4.	To establish a more definite and concrete sales policy.	S	9	13	- 4
5.	To develop a better community and help the general prosperity of the city.	P	15	4.5	+10.5
6.	To broaden its field of activity by making a wider variety of products.	M	7	12	- 5
7.	To make as fine a tire as any company can make and sell it a little cheaper.	M	5.5	4.5	+ 1
8.	To promote good labor relations and have satisfied workers.	P	4	2	+ 2
9.	To make a fair margin of profit.	F	3	14	-11
10.	To make a good and economical tire and sell it to private brand, mass distributors at a good price.	S	10	8	+ 2
11.	To increase the proportionate volume of Congo brand business handled.	S	11.5	15	- 3.5
12.	To develop an extensive, national advertising program.	S	17	17	0.0
13.	To keep the company growing and expanding.	GA	11.5	10	+ 1.5
14.	To expand the sales force in the field.	S	16	16	0.0
15.	To come out on top with a good quality tire and high production.	M	5.5	9	- 3.5
16.	To provide good working conditions and good living standards for employees.	P	8	1	+ 7
17.	To attempt to cut costs and expenses and put the company back in the profits brackets.	F	2	3	- 1

* Department Code: GA—General Administration; F—Finance; S—Sales; M—Manufacturing; P—Personnel. The department designated for each statement refers to the department of the company to which the statement best applied in the operational activities of the company.

departmental classification of each item, and also individual goals can be studied without regard for their departmental classification. Ideally it could be hoped that the rank order of the statements in terms of goal importance would be the same as the rank order of the statements in degree of achievement terms. This agreement would indicate a balance between these two important areas of executive activity and leadership. However, where differences are shown to exist, the opportunity is offered to top management to study the advisability of establishing policies which will restore a balance by concentrating executive attention and effort on those activities which it has been indicated were not in balance.

A study of the rank order of the statements by goal importance and degree of achievement, without consideration of the differences between the ranks, will provide information to top management which may be used in attempting to re-educate management in general or which will provide the basis for conference and consultation between management at various echelons regarding the proper place of various goals in the business and the extent to which these goals have been achieved.

For purposes of illustration here, those statements relating to the finance department (nos. 2, 9, and 17) and those relating to the personnel department (nos. 5, 8, and 16) will be examined. Statement 9 regarding a fair margin of profit

had a minus 11 difference, meaning that these executives believed the company had fallen far short of the great importance of this goal. In contrast to this, however, statement 2 "to make as much money as possible for the stockholders" received a plus two difference, indicating that these executives believed the company achievements had somewhat exceeded the goal importance of this activity. The remaining finance item with a minus one difference between ranks indicated that the company was working quite satisfactorily towards cutting expenses and putting the company back in the profits bracket, the successful completion of which should change the rank differences on statement 9.

When the statements relating to personnel activities were examined, it was noted that all three of the statements received a plus difference, meaning that in all of these activities the executives believed the company had achieved more than the goal importance of these personnel functions. The sensitivity of the Goal and Achievement Index form may be judged somewhat by the fact that, at the time the study was conducted, tire orders were falling off rapidly; the company was not operating at a profit; and it was the general executive opinion that the company could not re-establish itself on a profit basis in the highly competitive tire market until labor costs were reduced. It was also known that the union contract was about to come up for renegotiation and that in all probability there would be a demand for higher wages which would eliminate the possibility of decreasing labor costs.

Communication and Departmental Group Correlations

Table 2 gives the rank order correlations between goal and achievement ratings of the five departmental groups; and a composite of the president and general manager, vice president manufacturing, and treasurer (hereinafter referred to as PVPMT). The correlation of .47 for the total group actually was the rank order correlation between columns 2 and 3 in Table 1. Similar tables were prepared for the rankings of each departmental group, and the rank order correlations computed from them. The correlations indicated the extent

Table 2

Rank Order Correlations Between Goal and Achievement Composite Ratings of the 17 Items by Departmental Groups

	PVPMT	Mfg.	Sales
	.22	.49	.17
	Finance	General Adm.	Personnel
	.28	.66	.64

of agreement among executives in a given department on the relationship between the importance of company goals and the degree of achievement.²

Communication refers to how well the policies and thinking of top management have been carried through and absorbed by the remainder of the organization. If there were a condition of perfect communication within the company on the questions of goal importance and degree of achievement on the type of statement included in the Goal and Achievement Index form, then it could be expected that the ratings of each departmental group on all of the items would be identical with the ratings of the PVPMT, since they represent top management. If the ratings by departmental groups were the same, the rank order of each statement would be the same for each department, and consequently the rank order correlations of the various departments would be equal. These identical correlations might vary from minus 1.00 to plus 1.00, but as a measure of communication only, the size or the sign of the correlations bears no relationship to the effectiveness of communication. Instead, the effectiveness of communication can be measured only by observing the differences in the size of the correlations between various departmental groups and top management.

Therefore, the correlation of .66 for the general administration group expressed the highest group ratings toward the degree of achievement in terms of the goal importance of the various items. However, this correlation expressed the greatest disagreement with the opinion of the PVPMT and, therefore, indicated the least adequate communication.

² It should be noted that these correlations were descriptive of the relationships existing in this population of executives only, and cannot be considered as sampling statistics.

On the other hand, the correlation of .17 for the sales department, although it expressed the lowest departmental ratings toward the degree of achievement in terms of the goal importance of the items, was in closest agreement with the ratings of the PVPMT.

Analysis of Individual Statements

Although correlations and composite ratings are valuable for the general or over-all picture which they indicate, correlations obscure much detailed data which are available in the Goal and Achievement Index form and which can be analyzed by comparing individual statements in terms of departmental ratings, echelon levels, and individual executives. For the purpose of this article, one of the statements will be considered in terms of the differences in its ratings inter-departmentally.

On statement 10 (Table 1), there was a range of 11.0, the rank extending from 3.5 to 14.5. This statement was specifically important to a long-time company policy of selling tires under a private brand name (not Congo) to large retail outlet chain stores. The PVPMT gave this goal a 3.5 rank, indicating that this policy should be continued, and perhaps even emphasized. However, the sales department ranked this statement 14.5. This may have meant one of two things: (1) that the sales group was not aware of top management attitude toward this policy or (2) that the sales group was expressing its own opinion on the policy without regard to what its knowledge of top management attitude was. Under any possibility, statement 10 was classified as a sales statement regarding which the sales group deviated widely from top management attitude. The implications of this condition for the efficient operation of the company and the performance of the executive leadership within the company may be many.

When the statements were considered inter-departmentally in terms of achievement ranks, the highest range of 13.0 was for statement 1, classified as a manufacturing statement. The rankings extended from 3.0 to 16.0. The manufacturing group and the PVPMT gave this a 3.0 achievement rank, indicating their belief that the degree of achievement was well near the top of company achievements. How-

ever, the personnel group which reported to the vice president manufacturing ranked the statement 16.0, and the sales group ranked it next lowest with a 13.0.

Throughout these detailed comparisons, information was obtained which provided clues to the relationships existing inter-departmentally between executives in the business. One pattern which developed was a general tendency for executives in any given department to rank a statement which applied to their department in the higher ranks obtained by that particular statement. For example, a statement which applied to the sales department was likely to be ranked higher both for goal importance and for degree of achievement by the sales executives than by any other departmental group.

Rankings by Departmental Statements

In terms of efficient company and executive relationships as well as providing a measure of communication and knowledge of company goals and achievements, the differences between ranks assigned to statements grouped departmentally provided additional leading information. For example, when it was found that items which were given little or no goal importance were given high degree of achievement, or vice versa, then it appeared evident that there was a need for corrective activity in the particular aspect of company affairs covered by the item.

Table 3 presents the average rank differences between goal importance and degree of achievement rankings of the statements grouped according to the department to which they best applied. The rankings were the composite of each departmental group of executives. A plus value indicated that the degree of achievement for any given group of departmental statements as ranked by any given group of departmental executives was greater than the goal importance of the same group of departmental statements as ranked by the same group of departmental executives. A minus value, on the other hand, indicated that the degree of achievement for any given group of departmental statements as ranked by any given group of departmental executives had not reached the goal importance of the same

Table 3

Average Rank Differences Between Goal Importance and Degree of Achievement Rankings of Departmental Groups of Executives on Goal and Achievement Statements Classified into the Department to Which They Best Apply

Exec. Group	Genl. Adm. Statements		Finance Statements		Personnel Statements		Sales Statements		Manufacturing Statements	
	+	-	+	-	+	-	+	-	+	-
Total		1.7		3.3	6.5			1.1		.1
PVPMT		2.5		2.3	8.2			.4		2.4
Mfg.		4.5		1.0	6.5			1.2		.4
Sales		4.7		3.7	6.2		2.7			2.9
Personnel		.5		3.0	4.0			.6	0.0	0.0
Finance	.2			4.3	9.0			2.7		.5
Genl. Adm.		1.7		4.2	4.2		.4		.4	

group of departmental statements as ranked by the same group of departmental executives.

It will be noted that, with the exception of the personnel statements, there were only four cases of plus values. That is, the executives, regardless of department, believed that the company had over-achieved in those items which relate to personnel activity but in other activities there was almost consistent agreement on under-achievement. The personnel executives with an average plus difference in rankings of 4.0 expressed the smallest over-achievement value on personnel statements. On the horizontal rows, the manufacturing executives gave their lowest under-achievement value to manufacturing statements, and the sales executives gave a plus value or an over-achievement ranking on sales statements. On the other hand, however, the finance executives ranked their own departmental statements fifth or last. This attitude on their part may have been indicative of a somewhat cautious or pessimistic attitude on the part of those men who were guardians of the company purse, or it also may have been a direct reflection of the particularly acute financial situation in the company at that particular time.

Again in this phase of the study, there was evidence which suggested that the rankings given to goal and achievement statements were related to (1) the department to which the executive was assigned and (2) the type of work which the executive did.

Individual Goal and Achievement Scores

The individual goal and achievement scores and the three scores for each executive based

on the relationship between the goal and achievement scores are included in Table 4.³

For the G/A calculations a score of 1.00 indicated no difference between goal and achievement scores. In other words, the individual with a 1.00 G/A score ranked the 17 statements for degree of achievement of the company equal to the goal importance of the statements. Therefore, it follows that a G/A score of less than 1.00 indicated that the degree of achievement of the company had not equalled the goal importance of the statements, and a score of more than 1.00, that achievement had surpassed goal importance.

The range of the G/A scores was from .480 to 1.00. With the exception of the one case who ranked goal importance and achievement equal, all of the executives ranked achievement below goal importance.

The G minus A scores represented the algebraic difference of the goal score and the achievement score for the individual. All of the G minus A scores were minus, ranging from 0 to -28. This indicated that the degree of achievement had not reached goal importance, with the exception of the one score of 0 which indicated equality between achievement and goal importance. In this study it was somewhat simpler to work with the G minus A score since all of these scores were minuses or zero. If there were plus scores included, it

³ To reduce printing costs, Table 4 has been deposited with the American Documentation Institute and may be ordered as Document 2741 from American Documentation Institute, 1719 N Street, N.W., Washington 6, D. C., remitting \$0.50 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$0.50 for photocopies (6 × 8 inches) readable without optical aid.

would then simplify the study to work with the G/A score.

Both the G/A and G minus A scores can be used to measure the ratings of individuals toward company goals and their relationships to achievements of the company. However, to measure communication and knowledge of company goals and achievements as expressed by top management, some other score must be used. For this purpose, the algebraic difference between the individual executive's G minus A score and the G minus A score of the president and general manager was taken. All of these scores were plus values with the exception of three cases. This indicated that, with the three exceptions, all of the executives ranked the company's achievements in relation to goal importance greater than the president and general manager did. As a measure of communication, however, it was immaterial whether the individual's disagreement with the president and general manager was a plus or a minus score. The sign merely indicated the trend of the difference in terms of goal importance and achievement relationship, but a minus score had the same communication meaning as a plus score.

As another phase of the study, the G minus A scores and the G minus A deviations from the president and general manager were considered, arranged first by departmental groups of executives and then by echelon groups. The analyses of these results were characterized by two factors: (1) the executive contacts which the individual had within the business, and (2) the extent to which the work performed by the executive was related to the major over-all functions of the business. While this general pattern could not be supported and justified with each individual executive, the tendency of the pattern was strong enough to justify the recommendation that it be pursued further in other research.

Summary

The results of an exploratory study with a method for studying executive leadership called the Goal and Achievement Index have been presented, the data being based on a group of 23 executives in a tire and rubber manufacturing company. The data were ana-

lyzed for the total group of executives, for the executives grouped into the major departments of the business, and for individual executives.

On the basis of this exploratory study, it appears that the approach used in the Goal and Achievement Index offers promising possibilities for the study of executive function and leadership in business as it is related to (1) knowledge of company goals and achievements; (2) the relationship between goals and achievements; and (3) the effectiveness of the communication of this knowledge and relationship from top management through the supervisory levels of the business.

In the analysis of goal and achievement scores, two fundamental variables need to be particularly considered. The first has to do with the goal and achievement statements and the individual's evaluation of them at any given time. This variable is a constantly changing one, subject to variation in accordance with the conditions existing within a particular business at a particular moment.

The second variable is the relationship existing between the scores and ratings, inter-individually and inter-departmentally. It can be postulated that this relationship will remain relatively constant, unless there are changes in the personnel studied or unless the company attempts some type of a training program which will bring the executives, individually and departmentally, into greater agreement on and provide better understanding of company goals and achievements. Therefore, this factor is not subject to the same variability as the evaluation of individual goals and achievements since it is dependent upon the executives as individuals and not upon the constantly changing factors in company operations.

Received August 8, 1949.

References

1. Browne, C. G. An exploration into the use of certain methods for the study of executive function in business. Unpublished Ph.D. dissertation, The Ohio State University, 1948.
2. Browne, C. G. Study of executive leadership in business. I. The R, A, and D Scales. *J. appl. Psychol.*, 1949, 33, 521-526.
3. Browne, C. G. Study of executive leadership in business. II. Social group patterns. *J. appl. Psychol.*, 1950, 34, 12-15.

To What Extent Have the American People Accepted Socialism?

Henry C. Link and Albert D. Freiberg

The Psychological Corporation, New York City

In 1928, the national convention of the Socialist Party adopted a platform including, among others, the following domestic planks:

"1. Public ownership of all national resources and public utilities, including oil wells, coal mines, power plants, railroads, and telephone and telegraph facilities.

2. Governmental relief of unemployment by extension of all public works, Federal loans without interest to states and municipalities to promote public works, old age pensions, unemployment, health, and accident insurance, a shorter work day.

3. Acquisition of grain elevators, stock yards, and other distributing agencies by the government or by bonafide cooperating societies; cooperative purchasing, marketing, and credit agencies; government insurance against crop losses; flood control."¹

The intervening years have seen the enactment of some of the Socialist goals and the present administration has recommended legislation which would, in part, achieve others. The attitudes of present day citizens toward Socialism, their definition of socialistic aims, and their opinions of existing or proposed activities which might be regarded as implementation of socialistic philosophy constitute an interesting area for opinion research.

In a previous article,² we have reported the finding that 67% of an urban sample regarded Communism as a "dangerous thing in the United States" while 26% of the same group regarded Socialism as dangerous. Sixty-one percent of the respondents felt that Communism and Socialism were different.

In October, 1949, a random subsample of 1000 adults out of the 5000 interviewed in connection with the Psychological Barometer were asked a series of questions designed to yield a more detailed account of opinion on this topic.

¹ Socialist Party (U.S.) Socialist Campaign, 1928. As reported in the *WORLD ALMANAC* of 1929, the Republican National Convention of 1928, specifically endorsed none of these demands except for a mention of cooperative farm marketing agencies. The Democratic Convention favored government support (but not subsidy) of cooperative farm marketing agencies, public works to relieve unemployment, and flood control.

² Link, H. C., and Freiberg, A.D. The Psychological Barometer on Communism, Americanism and Socialism. *J. appl. Psychol.*, 1949, 33, 6-14.

Q. Are you for or against Socialism in this country?

Answers	%
Against Socialism	75.3
For Socialism	6.3
Don't Know	18.4

By the various socio-economic groups, these answers were as follows:

Answers	Upper	Upper Middle	Lower Middle	Lower
	%	%	%	%
Against Socialism	90	80	74	64
For Socialism	3	6	7	8
Don't Know	7	14	19	28

This was followed by:

Q. What does Socialism mean to you?

The 75% who were against Socialism and the 6% for Socialism described it as follows:

Answers	% Against	% For
Government ownership or control of industry, utilities, natural resources, health and welfare, medicine and doctors	26	30
Too much Government control; dictatorship; regimentation	18	
Distribution of wealth; fellow who produces gets no more than the one who doesn't	12	8
Destruction of capitalism; restricting or doing away with private enterprise and competition; disruption of our economic system; ruin of our country	8	2
Socialism is related to Communism, leads to Communism	9	
Equality for all; no racial discrimination; a hopeless crusade for equality	5	58
Socialism is like the English government of today	3	
Workmen's compensation; social security; the "Welfare State"	1	
Collectivism; collective Government; planned economy	1	6
Miscellaneous answers	9	10
Don't know	15	
Total Answers	809	72
Total Interviews (Base for per cents)	753	63

Evidently the majority who say they are against Socialism tend to identify Socialism with increasing Government control, operation and ownership.

Specific Issues

In order to bring the general issue of Socialism down to specific cases, the following series of questions was then asked of everyone in our 1000 interview sample.

- Q. Which of the following do you think are steps toward Socialism, and which are not?
- The minimum wage law where the Government sets a minimum wage below which no (interstate) business can pay?
 - The T.V.A. and other Government-owned flood control and electric power projects?
 - Payroll taxes or deductions from a person's pay to provide old age, unemployment and other benefits?
 - Food subsidies where the Government buys eggs, potatoes, and other farm products to keep their prices near parity?
 - Rent control where the Government decides how much rent can be asked for houses and apartments?
 - Government housing where the Government builds apartment houses which are rented to low income people at rents below actual cost?
 - Government rules which require that only union labor can be used on Government building?
 - Peace-time price control where the Government decides how much a business can ask for its products?

The responses of those 753 persons who previously had reported themselves as opposed to socialism are summarized at top of next column.

Conclusions

Although 75% of a sample of urban adults represent themselves as opposed to Socialism, within this 75% each of eight types of government activity which might logically be regarded as a specific instance of socialistic practice finds a considerable number who favor it and a

Items	Yes, It Is Socialistic	Favor This Measure
	%	%
a. Payroll taxes for old age and other benefits, etc.	27	81
b. T.V.A. and similar projects, etc.	37	66
c. The minimum wage law, etc.	38	62
d. Government housing, etc.	46	59
e. Rent control, etc.	50	43
f. Government rules requiring union labor on Government building, etc.	54	19
g. Food subsidies for farm products, etc.	55	24
h. Peace-time price control, etc.	67	18

number who do not regard the practice as "a step toward Socialism."

Four of these areas, payroll taxes for old age benefits, T.V.A. and flood control, minimum wage laws, and government housing find favor with a majority of those who oppose Socialism in general. Furthermore, in the case of each of these categories fewer than half of those against Socialism regard the measure as socialistic.

The more frequently a specific type of activity is regarded as Socialistic, the less frequently is it viewed with favor.

Some of the planks of the 1928 Socialist platform have been translated into practice. Several of these are not now regarded as "steps toward Socialism" and are favored by a majority of those adults who say they are against Socialism in the United States.

Even though Socialism may be a scare-word to 75% of the people, it does not follow that this fear will be translated into the fear of specific measures such as Government ownership of public utilities, government housing, "socialized" medicine and the like.

Received January 18, 1950.

Early publication.

The Selection of Patrolmen

Philip H. DuBois and Robert I. Watson

Washington University

Relatively little is known concerning selection of police officers by psychological methods. In a recent exhaustive review of the validity of commonly employed occupational tests, Ghiselli¹ reports briefly on previous work. Without either specifically identifying the particular tests utilized or the studies from which they came, he presents certain correlations with criteria of job proficiency. The median coefficients of validity with various kinds of tests were as follows: intelligence .28, immediate memory .40, arithmetic -.12, number comparison .25, name comparison .35, finger dexterity .20, and personality questionnaires .24. With the exception of intelligence in which eight coefficients were involved, most of these validities were based upon one or two correlations. The total combined N in all the studies numbered less than 500. The magnitude of the correlations, even in the absence of information about their reliability, shows that this field of personnel selection has certainly not yet been satisfactorily explored. With Ghiselli it is agreed that "about all that can be said with any degree of certainty is that intelligence tests give fair results and that personality measures would appear to be promising."²

The present study on the prediction of success of patrolmen in training and on the job was undertaken at the request of the St. Louis Board of Police Commissioners. In late 1947 an extensive reorganization of the Police Department was in progress. New equipment was secured, new office procedures were instituted, the crime detection laboratory was reorganized and expanded, a Police Academy for recruit and in-service training was organized, and plans were made for the introduction of psychological methods for the selection and evaluation of personnel. As a first step in the application of psychological methods to per-

sonnel problems, the authors were invited to participate in the selection of classes of probationary patrolmen.

The Sample

Table 1 presents information on the size and composition of the sample at various steps in the procedures. It will be noted that 1255 applications for appointment were processed. It was found, however, that only 512 of the men met the various requirements set up by the Board, as described in Table 1, as to education, age, height, weight, residence, citizenship, and satisfactory character. As many of the applications had been on file for some months, these 512 men were canvassed as to their continued interest in being appointed to the Force. The 312 still interested were invited to take the selection tests, with 253 actually appearing for the examination.

Of these, 180 were found to meet the selection criteria set up for the psychological test battery shortly to be described. Twelve men were eliminated by failure to pass the physical

Table 1
Derivation of the Sample Studied

	N
Applications for appointment	1255
Applicants meeting established standards (Height 5'9" to 6'4"; weight 150 to 210 lbs.; age 22 to 31 years with year-for-year waiver for military service; residence 2 years in city; citizenship; education 10th grade or equivalent; good character as shown by police investigation)	512
Applicants requesting examination	312
Applicants appearing for examination	253
Examinees meeting cutting scores (AGCT score of 100 or more, CWF-2 of 9 or less)	180
Passed physical examination	168
Passed by Screening Board	139
Entered Classes 1 and 2	139
Sample used in present study	129

¹ Ghiselli, E. E. The validity of commonly employed occupational tests. *Univ. Calif. Publ. Psychol.*, 1949, 5, No. 9, 253-288.

² *Op. cit.*, p. 269.

examination, leaving 168 who were interviewed by a Departmental Screening Board. The Screening Board consisting of five members of the Department, including the Chief of Police, rated each man on five traits: appearance, manner, speech, adaptability, and general impression, giving scores of 1 for the lowest and 5 as the highest score in each of the characteristics. A score of 125 was perfect and a score of 75 or above was considered acceptable. Twenty-nine candidates were eliminated on the basis of the screening board rating. The remaining 139 were appointed as probationary patrolmen, entering the Police Academy in two successive eight-week classes, which graduated in June and August, 1948. The present study is concerned with 129 of these 139 cases. Of the ten cases not being reported, two resigned while in good standing at the Academy, five failed in the curriculum, while in the three other instances data were incomplete for one reason or another. The number of failures in training was not large enough for consideration in this paper, although the problem is scheduled for study at a later date.

Tests Administered

The battery of psychological selection and experimental tests used in this study is listed in Table 2. Before turning to the results each instrument will be described.

The St. Louis Police Aptitude Test was developed by the writers in consultation with the departmental staff. All material was prepared in the police context, so that it had high face validity. Realism was obtained by using police photographs and excerpts from police manuals. The items were screened by experienced police officers who decided whether or not they were plausible in a police setting. In its final form the test consisted of 90 multiple choice items divided into five sections. The first section involved 15 memory items. They were of two sorts: memory for photographs and names, and memory of a scene of a crime and of a scene of an accident. A touch of verisimilitude was added by using the photographs of recent criminal suspects. The names were selected at random. The accident and crime scenes were posed and involved in each instance a variety of clues. Typical of

the questions asked was one involving the locus of the crime. The second section of 15 items required the applicant to select the one misspelled word from a list of five or to select the one correctly spelled from a list of words otherwise incorrectly spelled. Typical words included alias, waiver, accomplice and felony. Three reading selections made up the third section. These were reports of a fire, a robbery, and a selection from the police manual concerning rules in regard to the making of arrests. Each was followed by five multiple choice reading comprehension items. The next section consisted of 30 items concerned with general information and judgment related to police work. Police practice, causes of crime, the purpose of teaching first aid to policemen and the reason automobiles need to be inspected regularly were typical of the bases for these questions. Care was taken that specialized knowledge generally known only to a person with police experience was not required. The last section involved 15 arithmetical problems couched in a police department setting. The test as a whole required 90 minutes, including directions. Originally it was proposed to use a cutting score with this test, but since the correlation with the AGCT was higher than anticipated it was used only experimentally with the first two classes.

The Army General Classification Test, First Civilian Edition, was actually used for selection. An AGCT score of 100, corresponding to a percentile rank of 48, was the minimum acceptable. The two classes had a mean AGCT score of 118 or a percentile rank of 80 in the general population. This score stands at the 75th percentile of the distribution reported for policemen.

Candidates were asked to write a short composition entitled "Why I Wish to Become a Policeman." This essay was graded for handwriting, using the Thorndike Handwriting Scale, and for coherence in English expression, using a five-point scale. No candidates were actually eliminated on this test but the lowest rating essays of men who were otherwise qualified were scanned by the Chief of Police for acceptability. The final test included in the original screening battery was the Cornell Word Form—2, a controlled association test which has been used to detect neurotic ten-

dencies. Scores of ten or more were considered disqualifying.

In the course of the eight weeks training at the Police Academy it was possible to administer for experimental purposes a further series of tests. These included the Figure Matching Test, a speeded perceptual measure involving matching the silhouette of an object or a meaningless shape with five similar figures, one of which is the exact duplicate; the Bennett Mechanical Comprehension Test—Form BB; the Minnesota Paper Form Board MB; the

Object-Aperture Test, an instrument requiring visualization in three dimensions and yielding a speed score and a power score; the Strong Vocational Interest Blank, which was completely scored but which has been validated for only three scales—police interest, interest maturity, and occupational level; and, finally, the Rosenzweig Picture Frustration Study, developed as a measure of the direction of aggression and the type of frustration reaction.

The Strong and the Rosenzweig measures were introduced to study the validities of

Table 2

Validities of Psychological Tests and Screening Board Rating for Prediction of Success as Probationary Patrolman (N = 129). Classes of June, 1948, and August, 1948, of Police Academy, Department of Police, City of St. Louis

	Academy Grade June Class (N = 72)	Academy Grade August Class (N = 57)	Achieve- ment Test (N = 129)	Marksman- ship (N = 129)	Service Rating (N = 129)
Police Aptitude Test—Total	.39**	.50**	.47**	.08	.03
Aussage (15 items)	.40**	.20	.25**	.20*	.08
Spelling (15 items)	.44**	.42**	.38**	-.10	-.03
Reading (15 items)	.18	.40**	.41**	.11	-.03
Information and Judgment (30 items)	.32**	.37**	.36**	-.01	.12
Arithmetic (15 items)	.31**	.30*	.42**	.14	-.07
Army Gen. Classif. Test—Total	.54**	.50**	.47**	.15	.10
Verbal Items	.53**	.53**	.55**	.02	.03
Numerical Items	.41**	.43**	.36**	.11	.02
Blocks	.37**	.30*	.23**	.16	.14
Handwriting	-.03	.04	.06	-.06	.10
Composition	.23*	.30*	.26**	-.06	.08
Cornell Word Form 2	-.14	-.18	-.22*	-.09	.06
Figure Matching (speeded perceptual)	.25*	.29*	.23**	.19*	.03
Bennett Mech. Comp. BB	.28*	.29*	.20*	.27**	.10
Minn. Paper Form Board MB	.38**	.29*	.30**	.26**	.04
Object-Aperture Test—Speed	.18	.14	.09	.07	.10
Object-Aperture Test—Power	.28*	.29*	.17*	.17*	.12
Strong Voc. Int.—Police Interest	-.09	-.12	-.24**	.15	-.01
Strong Voc. Int.—Interest Maturity	.14	-.17	-.05	-.01	-.04
Strong Voc. Int.—Occupational Level	.06	.19	.30**	.01	.03
Rosenzweig Picture Frustration					
Obstacle Dominance	-.03	-.09	-.11	.12	.08
Ego Defensiveness	.06	-.17	-.01	-.17*	.12
Need Persistence	-.04	.16	.08	.03	-.10
Impunitiveness	.24*	.04	.10	.13	-.05
Intropunitiveness	-.16	-.04	-.04	.03	.12
Extrapunitiveness	-.11	-.02	-.03	-.14	.10
% Group Conformity Rating	-.01	.02	-.17*	.13	.04
Screening Board Rating	.01	.11	-.01	.06	-.03

** Significant at 1% level.

* Significant at 5% level.

certain interest and personality characteristics for police work. The other experimental tests were non-verbal instruments which were considered to be possibilities for supplementing the selection tests, success in which depended largely upon verbal ability.

Validities

Four criteria of validity are used in this study: the final grade in the Police Academy, treated separately for the two classes; an achievement test based on Perkins' "Elements of Police Science," which was required reading after completion of the course and on which the patrolmen took an examination approximately 10 weeks after graduation; marksmanship during the academic training; and service rating after 10 weeks on duty. The service rating was based upon an 11-trait rating scale with five descriptive steps for each trait. The traits were work attitude; loyalty, interest and enthusiasm; judgment; report writing; investigative ability; alertness; bearing and demeanor; speech; appearance; contacts with the public; and usefulness to the service. Ratings were made by the sergeant who directly supervised the work of the patrolman and were approved by superior officers in the chain of command. Only the validities against the over-all service rating are reported at this time.

In Table 2 validities are reported for all predictor variables against these four criteria. It will be noted that for the criteria of academy grades and the achievement test, the verbal and numerical sections of the Police Aptitude Test and Army General Classification Test have the highest validities, closely followed by the Composition Test and the non-verbal experimental tests. For 72 cases reported in the June Class there are ten correlations significant at the 1% level and five more at the 5% level. For the August class of 57 cases, there are seven correlations significant at the 1% level and seven more significant at the 5% level. Against the Achievement Test, where the N is 129, there are 15 correlations significant at the 1% level and four more significant at the 5% level. Against the criterion of marksmanship there are only two correlations significant at the 1% level and four at the 5% level. It will be noted that except for one of the Rosenzweig

categories, all the correlations which are significant against the marksmanship criterion are non-verbal aptitude measures. None of the individual correlations with service rating is statistically significant at the 5% level.

The negative validities of the Police Interest Scale of the Strong Vocational Interest Blank against all criteria except marksmanship are perhaps surprising. However, the academy grade and the Achievement Test would probably be somewhat artificial criteria of performance for policemen on whom the scale was developed. A tabulation of the Police Interest Scores of the men in the first class showed 86% of them A or B+, with no score below B-. It is possible that in a greater range of interest scores, the test would have been predictive.

The complete matrix of approximately 1000 intercorrelations of all the predictor variables and the several criteria of success has been computed. The intercorrelations of 22 selected predictor variables are presented in Table 3.³ The Police Aptitude Test has a correlation of .53 with the Army General Classification Test in this restricted range. The intercorrelations of the non-verbal experimental tests ranged from .20 to .55, indicating that there is considerable over-lap among them.

In Table 4, Beta weights, together with multiple correlations, are presented for the prediction of several criteria of success. For academic grades and achievement tests, the Wherry-Doolittle test selection method was employed. For marksmanship and service ratings ordinary regression weights were computed using the variables which appeared to be the most promising.

It is immediately apparent that the combinations of tests useful in prediction differ considerably from criterion to criterion. The Police Aptitude Test, the Army General Classification Test and at least one of the non-verbal tests (Minnesota Paper Form Board or Bennett Mechanical Comprehension) appear in the regression equations predicting the

³ To save space, Table 3 has been deposited with the American Documentation Institute and may be ordered as Document 2696 from American Documentation Institute, 1719 N Street, N.W., Washington 6, D. C., remitting \$.50 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$.50 for photocopies (6 X 8 inches) readable without optical aid.

Table 4

Beta Weights and Multiple R's in Predicting Several Criteria of Success as Probationary Patrolman, St. Louis Police Department

	Academic Grade Class 1 (N = 72)	Academic Grade Class 2 (N = 57)	Achieve- ment Test (N = 129)	Marksman- ship (N = 129)	Service Rating (N = 129)
Police Aptitude Test—Total	.11	.28	.28	.12	
Aussage					.08
Information and Judgment					
Army Gen. Classif. Test—Total	.41	.28	.25		.11
AGCT—Blocks					
Composition		.13			
Figure Matching				.17	
Bennett Mech. Comp.—BB		.12		.15	
Minn. Paper Form Board MB	.20		.11		.13
Object-Aperture (Power)			.23		
Strong—Occupation Level					
Rosenzweig					.20
Ego Defensiveness					.18
Intropunitiveness					
Multiple R	.58**	.60**	.59**	.33**	.29*

** Significant at 1% level.

* Significant at 5% level.

"academic" criteria, i.e., academic grade in Class 1 and Class 2, and the achievement test. The composition test carries a weight in Class 2 and the Strong-Occupational Level carries a weight for the achievement test.

Marksmanship can be predicted to some degree from tests of a non-verbal sort, involving perception, space relations, and mechanical comprehension.

Service rating, which is predicted least accurately, depends to a considerable degree upon personality predictors, with some weight being carried by the tests of spatial relations and information and judgment. For service rating the multiple of .29 is significant at the 5% level of confidence, while all other multiple R's are significant at the 1% level. It may be that tests that will predict the service rating more adequately will have to be devised.

A third class, numbering 36 cases and selected in part by means of a composite score based on the Police Aptitude Test, the AGCT, the Composition Test and the Cornell Word Form, recently completed a 12-weeks training course. For this group the correlation between a weighted score based upon the re-

gression weights obtained with Class 2 and academic grade was .62, or slightly better than anticipated. Of the 12 men who were in the top third according to this composite score, 83% had final grades above the median for the class, compared with 41% in the middle third and only 25% in the lower third. Since in Class 3 the validity of the Police Aptitude Test was .51 and of the AGCT .55, it would appear that the battery as a whole is somewhat more predictive in police training than a mental alertness test alone would be.

Plans for Further Research

Up to the present time only the first edition of the Police Aptitude Test has been used. Prior to its administration in connection with Classes 1 and 2, none of the material had been tried out in any way. An item analysis of this test has been completed which is being used as a guide in preparing the next edition. As a number of the items have been found to be unsuitable, it should be possible to improve the validity of the instrument.

Plans for further research include validating

the original battery against new criteria as they become available, such as service ratings after a year and two years of service, remaining on the force over a period of years and special reports of meritorious service. In the first three classes, the experimental tests administered were largely of the non-verbal sort, and from that study we can add two or three tests to the selection battery. Other types of tests are scheduled for experimental try-out from time to time, so that the selection battery can be modified as further validity information becomes available.

Summary

1. A battery of psychological selection and experimental tests was administered to two entering classes ($N=72$ and $N=57$, respectively) of probationary patrolmen at the St. Louis Police Academy. A third class ($N=36$) was used to test the predictive efficacy of the battery. Criterion variables were final grade in the Police Academy, an achievement test on knowledge of police practices, marksmanship and ratings by superior officers.

2. A specially constructed Police Aptitude Test and the Army General Classification Test were consistently good predictors of academic performance and of the achievement test score. At least one non-verbal test, such as the Bennett Mechanical Comprehension Test, or the Minnesota Paper Form Board, is needed as a supplement in order to achieve a multiple coefficient of validity in the neighborhood of .60.

3. The best predictors of marksmanship are non-verbal aptitude tests, a combination of three of which yield a multiple coefficient of validity of .33.

4. None of the tests explored has a significant correlation with rating on the job by superior officers. A combination of three non-verbal tests and two variables of the Rosenzweig Picture Frustration Study yields a multiple coefficient of validity of .29 which is significant at the 5% level of confidence.

5. In the third class ($N=36$) a predictive score, based upon the regression weights developed with Class 2, had a correlation of .62 with final grade in the Police Academy.

Received August 3, 1949.

Group Differences in Performance on the Meier Art Test

E. Terry Prothro

University of Tennessee

and

Harold T. Perry

Louisiana State University

The Meier Art Judgment Test attempts to measure an ability which is subject to considerable development through learning and experience. In this paper we have compared the scores of several groups of subjects in an effort to determine whether membership in those groups is associated with differences in test performance. If it is, then such information would be of importance to those counselors who utilize the test and to those students of aesthetics who wish to understand the nature and genesis of art judgment.

Subjects

A total of 410 subjects were used in this study. Of these, 109 were from a state university for whites, 100 were from a state university for Negroes, 100 were from the tenth and eleventh grades of a city high school for whites, and 101 were from the tenth and eleventh grades of a city high school for Negroes. All four schools are located in the same city in the state of Louisiana. Of the total number of subjects, 223 were males and 187 were females.

Procedure

Selection of subjects was by school classes. Permission was obtained from the school authorities to test all of the members of a given class. Neither authorities nor students were given advance information on the exact nature of the study. The revised Meier Art Judgment Test booklets and forms were placed on the desks of all students in the class. Each class consisted of approximately 30 students.

Because the test is self-administering, few verbal instructions were given. It was explained that the test was not a part of the

school curriculum, and that performance on it would not affect school standing in any way. Only two persons objected to taking the test; both were excused. When the subjects filled out the personal data sheets, they were specifically instructed to indicate the number of craftsmen in their ancestry by filling in the blanks at the bottom of that sheet. As far as could be determined by the experimenters, test conditions were about the same for all subjects.

Results

Examination of the personal data sheets revealed that a majority of the subjects had received some art training in elementary school, some had studied art in high school, but very few had had any training beyond that.

The results are summarized in Table 1. It can be seen that there are significant differences between mean scores of every pair of groups except the male-female groups. The similarity of mean scores of the male students and the female students was present in each of the four schools. In the two Negro schools, the males did slightly better; in the two white schools, the females did slightly better. None of these differences was significant at the five per cent level. The similarity of the sex groups on performance on the Meier Test is especially notable in view of the fact that Carroll and Eurich¹ found that women did better than men on both the McAdory and the Meier-Seashore tests.

The superiority of university students to high school students and of white students to Negro students may be attributable to differences in the socio-economic status of the

¹ Carroll, H. A., and Eurich, A. C. Group differences in art judgment. *Sch. & Soc.*, 1931, 34, 204-206.

Table 1
Comparison of Mean Scores on the Meier Art Test

Group	N	Mean	S.D.	Diff.	S.E. of Diff.	t
White University	109	96.6	9.7			
White High School	100	89.7	10.1	6.9	1.38	5.06**
Negro University	100	78.3	11.2	11.4	1.52	7.50**
Negro High School	101	74.7	10.5	3.6	1.54	2.34*
University Students	209	87.8	13.9			
High School Students	201	82.1	12.8	5.7	1.32	4.33**
White Students	209	93.3	10.5			
Negro Students	201	76.5	11.0	16.8	1.12	14.98**
Male Students	223	85.6	13.3			
Female Students	187	84.3	14.1	1.3	1.36	0.94
Students with Craftsmen in Ancestry	135	87.2	13.9			
Students with No Craftsmen in Ancestry	275	84.0	13.4	3.2	1.45	2.23*

* t-score is significant at the five per cent level.

** t-score is significant at the one per cent level.

groups. Rodgers² found that better environments surrounded children selected as more competent artistically. Meier³ has pointed out that some aspects of artistic ability are subject to considerable development through experience, and we believe that the differences in daily experiences of members of different socio-economic groups may be the basis for the differences in their scores.

Our results confirm Meier's thesis that there is some relationship between "craftsman heredity" and aesthetic judgment. The differences between persons with craftsmen in their ancestry and those with no known craftsmen in their ancestry are fairly small, but they are significant at the five per cent level. It is possible that the ancestry factor is less significant for unselected school populations than for

populations of art students. Whether the relationship is a function of selective recall about ancestry, of differences in home environment, or of some "innate neuro-physical constitution" cannot be determined from the information now available.

Summary

The revised Meier Art Judgment Test was given to 410 high school and college students in Louisiana. Comparison of mean scores revealed that the white students did significantly better than the Negro students and that college students did significantly better than high school students. It was suggested that these differences may be associated with differences in socio-economic status of the several groups. Those persons who had craftsmen in their ancestry did somewhat better than those who did not have craftsmen in their ancestry. There were no significant sex differences.

Received June 10, 1949.

² Rodgers, F. Variation in the aesthetic environment of artistic and non-artistic children. *Psychol. Monogr.*, 1933, 45, No. 1, 95-107.

³ Meier, N. C. Factors in artistic aptitude: final summary of a ten-year study of special ability. *Psychol. Monogr.*, 1939, 51, No. 5, 140-158.

Use of an Objectivity Key on a Short Industrial Personality Questionnaire

Harold F. Rothe

Stevenson, Jordan and Harrison, Inc., Chicago, Illinois

Industrial and personnel psychologists have long been concerned with the problem of possible "faking" by respondents to personality questionnaires. Bonnardel (1), in a discussion of industrial psychology in France, wrote that such questionnaires are little used because it is believed that candidates for jobs give the responses most favorable to their obtaining the jobs. Giese and Christy, in a study reported by Tiffin (4), found that the Humm-Wadsworth Temperament Scale could be "faked" by college students to show higher normal (N) scores and less of undesirable characteristics. Longstaff (3) has recently shown that the Strong Vocational Interest Blank and the Kuder Preference Record could both be faked. He suggests that a key like the L or K keys of the Minnesota Multiphasic Personality Inventory be incorporated in industrial questionnaires, not because faking is known to exist, but because it might exist.

The purpose of the present paper is to report some experience in the use of a key similar to the MMPI L key on a short personality questionnaire that has been used in industrial practice for the past three years. This form has been developed for the use of the members of the psychological staff of a management consulting company and is not available for general distribution. Despite this restriction of the particular form, the experiences gained with it are believed to be helpful to other psychologists, and hence this paper is presented.

Description of the Form

The form used here consists of fifty items to be answered by a YES or NO. It is self-administering, has no time limit, and can be used for individuals or groups. All items are stated in a positive manner and most of them refer to behavior rather than to feelings, attitudes, or interests. The items are scored on a

number of keys that are industrially significant. One key is the so-called Objectivity key which has been patterned after the L scale of the MMPI. The other keys are based upon items which have been found valid, to some extent, on other published keys, although in each instance the wording has been slightly changed to avoid copyright infringement. Several keys have only three or four items, and others run as high as ten or twelve. For statistical reasons, only the longer keys have been analyzed and are mentioned in this paper.

In one very important respect this questionnaire differs from most standardized personality questionnaires. This form is primarily an interview aid. It is used by trained persons in conjunction with a personal history of each respondent, various tests, and an interview. Being used in this manner, the extreme shortness (50 items) is an advantage rather than a handicap. The interviewer can look over each statement and each answer, and can interview about each item if he so desires. At the same time the form can be scored to reveal patterns, and the interviewer can also use this pattern information as a part of his interview.

The longer keys are Emotionality, Social Dominance, Drive, Extraversion, and Objectivity. Since Social Dominance and Extraversion are highly correlated with each other (in this unrevised form), only the Social Dominance key is discussed here. The Emotionality key has been found to be correlated with neuroticism and may be considered as a measure of that. The Drive key measures the aggressiveness with which the respondent strives to reach various goals. It is planned to present validation studies for some of these keys in subsequent papers and accordingly that problem is not discussed further here.

Reliability was determined by testing and retesting a class of 44 college students with a

two-week intervening period.¹ The reliability coefficients (or better, the consistency coefficients) were .84 for the Objectivity, Emotional, and Social Dominance keys, and .69 for the Drive key. These compare very favorably with the test-retest reliabilities of .71 to .83 reported for the MMPI (2, p. 3). The consistencies reported in this paper are remarkably high considering the small number of items on each scale. The high degree of consistency is undoubtedly partly a function of the well-controlled conditions under which the test-retest was made.

The Objectivity key, as stated before, is closely patterned after six of the items on the L scale of the MMPI. It would perhaps be more accurate to refer to these items as L items so that the source of the origination would be more clearly recognized and credited. The name was changed for practical purposes. Since the profile obtained from this form is frequently discussed with employers or prospective employers, it may be an injustice to state that a respondent is "lying" or "faking." Even though this may be true, insofar as this questionnaire is concerned, the fact may well be that the respondent is basically an honest person. It might be doing him a great injustice to let someone make conclusions about the extent to which he "lies" when that conclusion would be based upon six items on a questionnaire.

Relation of Objectivity to Emotionality

Since the Emotionality key has been found to discriminate neurotics from the general industrial population with a fairly high degree of accuracy, the primary analysis was made upon the relation of the Emotionality scores to the Objectivity scores. The work of Giese and Christy led to the belief that less objective persons might show lower Emotionality scores than highly objective persons. Conversely, highly objective persons might be expected to show high Emotionality scores. This assumption is based upon the assumption that the Humm-Wadsworth N score is reasonably valid. It is also based upon the common-

sense observation that if someone is going to fake a questionnaire he will fake items to make him look better adjusted or better qualified for the job, as he interprets these items.

A sample of 525 persons was selected. The selection was made on the basis of their Objectivity scores; 144 persons scoring 0 or 1, 237 persons scoring 2 or 3, and 144 persons scoring 4, 5, or 6. The sample was selected from master data sheets on which are recorded scores of all persons taking this and other tests. These persons were all industrial workers in various job classifications. For the most part they were supervisors, or supervisory candidates, in both shop and office jobs. Beginning with the most recent entry, the first 144 persons who scored 0 or 1 in Objectivity were located, and their Emotionality scores were plotted. In a similar manner the other two groups were located and plotted.

The relation between Emotionality and Objectivity could be determined by correlational techniques, but have not. This is because of the writer's belief that a table of frequency distributions would show the same thing graphically and in a more meaningful manner. Table 1 contains the distribution of Emotion-

Table 1
Distributions of Emotionality Scores Made by
525 Persons Obtaining Different
Objectivity Scores

Emotion- ality Score	Number of Persons with Objectivity Scores of		
	0 or 1	2 or 3	4, 5 or 6
0	5	2	0
1	17	11	4
2	25	20	5
3	29	41	17
4	24	38	15
5	26	37	21
6	10	34	20
7	2	24	20
8	4	20	18
9	2	8	8
10		2	10
11			4
12			1
13			1
14			
Mean	3.5	4.8	5.7
S.D.	1.9	2.1	3.4

¹ The writer wishes to thank Robert C. McWilliams of Northwestern University for conducting this experiment.

Table 2

Distributions of Social Dominance Scores Made by 525 Persons Obtaining Different Objectivity Scores

Social Dominance Score	Number of Persons with Objectivity Scores of		
	0 or 1	2 or 3	4, 5 or 6
0	1	0	0
1	0	1	0
2	2	3	1
3	3	10	13
4	7	18	11
5	14	39	26
6	36	40	37
7	39	68	29
8	38	45	22
9	4	13	5
10	0	0	0
Mean	6.5	6.3	6.0
S.D.	1.5	1.6	1.6

ality scores for these three groups (High, Low, and Medium) selected on the basis of their Objectivity scores.

Table 1 shows that persons with high Objectivity scores tend to have high Emotionality scores, and vice versa. The critical ratio between low and medium is 2.4; between medium and high is 1.3; and between low and high is 3.3. The first two of these differences may be said to be of little statistical significance.

It is perhaps more meaningful, from an interviewer's point of view, to look at the distributions. A respondent who scores 7 on the Emotionality key may be considered as probably emotional or neurotic if his Objectivity score is 0 or 1. But if his Objectivity score is 5, that person would be practically average in Emotionality, and the interviewer would not expect to find symptoms of neuroticism. The interviewer could profitably spend more time interviewing in other areas with this highly objective respondent.

It is possible that the inverse relationship of Objectivity to Emotionality that has been found here is independent of "faking," i.e., that non-objective persons are actually less emotional or less neurotic than objective persons. Data exist, and it is hoped to publish them shortly, to rule out this possibility.

Relation of Objectivity to Social Dominance and to Drive

There is, as shown in Table 1, a relation between Objectivity scores and Emotionality scores, for this industrial sample. This raises the question of the extent to which scores on other scales vary with Objectivity scores for these persons. Two other long scales of this questionnaire have been selected for further analysis.

Table 2 contains the distributions of the scores of these same three groups on the Social Dominance scale. There is very little difference among the Low, Medium, and High Objectivity groups in terms of their Social Dominance. The critical ratios are: .44 between Low and Medium; .70 between Medium and High; and 1.08 between Low and High. None of these differences is significant.

Table 3 contains the distributions of the same three groups on the Drive scale. Again the differences between the groups are insignificant with critical ratios of .75 between Low and Medium; .68 between Medium and High; and 1.30 between Low and High.

The results presented in these three tables are based upon an industrial sample composed of adults, most of whom have had supervisory experience of one kind or another. The sample is uncontrolled in terms of intelligence and it is

Table 3

Distributions of Drive Scores Made by 525 Persons Obtaining Different Objectivity Scores

Drive Score	Number of Persons with Objectivity Scores of		
	0 or 1	2 or 3	4, 5 or 6
0	0	0	0
1	0	1	1
2	1	1	1
3	6	10	5
4	20	25	8
5	38	53	20
6	35	61	42
7	34	40	36
8	8	33	23
9	2	12	6
10	0	1	2
Mean	5.6	6.0	6.3
S.D.	1.3	1.2	1.6

assumed that there are no differences among the three groups in that respect. It does not necessarily follow that similar results would be obtained from samples of a different population.

To be more explicit, it is possible that other groups might show a variation in Social Dominance or Drive scores, depending upon Objectivity scores. It is reasonable to assume that the persons analyzed here, having had supervisory experience, might truly rank fairly high in both Social Dominance and Drive and hence not attempt to "fake" higher scores in those respects. It is also possible, however, that these persons could not "fake" either of those keys, even though they tried to do so. This possibility seems less likely than does the previous one.

Summary and Discussion

Data have been presented to show how the Emotionality scores of persons on a short industrial personality questionnaire may vary with the Objectivity, or L scale, scores of those persons. Scores on Social Dominance and Drive scales do not vary significantly with Objectivity scores. This means that persons may "fake" a better adjustment than they actually have; it also means that the interviewer can adjust for the faking by interpreting an Emotionality score on norms based upon persons of similar Objectivity scores.

This paper accordingly illustrates a method for correcting for faking on short industrial personality questionnaires. This is essentially the same technique as that of the L scale in the MMPI, and the actual items are similar to MMPI items (2). There is, however, one interesting difference in terms of interpretation.

One aspect of industrial psychology is the testing and interviewing of the key personnel within an organization to determine the extent to which those persons are adequately qualified for their positions. It is reasonable to assume that these persons will "try their best" on the various tests and questionnaires. Indeed, it

is the interviewer's responsibility to see that they do. Under these conditions it is only to be expected that a small degree of "faking" or of "stretching the point" will prevail. The interviewer expects this and includes it as a part of his frame of reference.

Occasionally, however, an extremely objective, or perhaps an extremely naive, person is encountered who is frank about himself to the point where he penalizes himself on a standardized questionnaire. Such a person might admit that he loses his temper, for example, even though he realizes this statement may be weighed against him.

Probably the chief use of an Objectivity key such as the one described in this paper is to protect such persons. That is, the problem is not one of locating "fakers" and possibly adjusting their results downward. Rather, the problem is one of protecting the "overly honest" persons and adjusting their scores upwards, so that they may be compared favorably with their more average fellows. An example of this has previously been given: A high Emotionality, or neurotic, score may not be serious if the respondent is highly objective; the same score may be very extreme if the respondent is non-objective. The interviewer must include the concept of objectivity in his interpretation to obtain a correct description of the respondent. One method for doing so is to maintain different norms for different objectivity groups, as has been described here.

Received August 9, 1949.

References

1. Bonnardel, R. Industrial psychology in France. *Pers. Psychol.*, 1949, 2, 47-68.
2. Hathaway, S. R., and McKinley, J. C. *Manual for the Minnesota Multiphasic Personality Inventory*. New York: The Psychological Corporation, 1943.
3. Longstaff, H. P. Fakability of the Strong Interest Blank and the Kuder Preference Record. *J. appl. Psychol.*, 1948, 32, 360-369.
4. Tiffin, J. *Industrial psychology*. New York: Prentice-Hall, Inc., 1947, rev. ed., pp. 170-171.

The Application of Weber's Law to Job Evaluation Estimates

Edward N. Hay

Edward N. Hay and Associates, Inc., Philadelphia, Pa.

One method of job evaluation which is gradually coming into more general use is the method known as factor comparison, devised by Eugene J. Benge.¹ Benge's method is similar to other methods of job evaluation such as the predetermined point plan in that the judgments are expressed in points, and both methods evaluate *factors* of the job under study rather than the job as a whole. The scales used by the two methods are quite different. In predetermined point plans certain qualities are identified, on the basis of a study of the jobs, as necessary for their performance. Scales are then constructed and given point values *a priori* to measure these requirements. Factor comparison scales, on the other hand, are constructed by dividing the salaries of existing jobs among the factors. Nowadays, however, the per cent method² of creating factor comparison key scales is frequently used. This method expresses as percentages, (1) the magnitude of each factor of a job in relation to the other factors in that job, and (2) the relative magnitude of a factor in one job to the same factor in the other key jobs.

Job evaluation by any method is dependent on the judgment of the evaluators. In point methods much, if not most, of the judgment is in the factors; their selection, definition, pointing, and weighing. In factor comparison the judgment is less in these things, and more in the comparison of factors of jobs being evaluated. The evaluators compare one job with another in respect of each factor rather than match a job against a description on a scale. The jobs are familiar to all members of the evaluation committee because the original

scale of jobs is created by the committee itself. Point methods tend to have steps of irregular spacing, while factor comparison evaluation scales have intervals that are equal in "sensed units of difference."

Not much has been written about the theory underlying the use of this method. It has been found, however, that the judgments expressed in factor comparison job rating seem to behave in accordance with Weber's Law.³ By way of illustration: an increase in job difficulty at a high level calls for a larger salary "raise" than at a low level. For example, at the lower end of the scale two jobs of just noticeable difference in difficulty might have salaries of \$100 and \$120 a month respectively. At a higher level two jobs of just noticeable difference in difficulty might be paid \$1000 and \$1200 respectively, but certainly not \$1000 and \$1020. So arithmetic intervals in job difficulty are matched by logarithmic increments in salary.

One of the steps in the factor comparison method calls for assigning a percentage value to each of the factors in a job (factors which are common to all jobs) in such a way that the total of all the factor values equals 100%. For example, the job of machine bookkeeper was selected as a "key job," and the committee made the following assignment of percentage values to the five factors as representing in its judgment the relative importance of those factors in that job: Mental Requirements, 18; Skill Requirements, 27; Physical Requirements, 17; Responsibility, 27; and Working Conditions, 11.

The factor scales used by the author are constructed with steps of 15 per cent increment and uniform differences in job difficulty. Table 1 shows steps for the factor of skill, and is representative of other factor scales created by evaluation committees.

Additional jobs that are to be evaluated are

¹ Described in *Industrial Relations Magazine*, Feb., Mar., and Apr., 1932 (now out of print), under the title "Gauging the job's worth." The mechanics of the method are also described in Benge, Burke, and Hay, *Manual of job evaluation*. Harper and Bros., 1941, in chap. IV, pp. 41-56, which is a reprint of an article by E. N. Hay, Arranging the right pay, *Personnel J.*, 1939, 17, 1-7.

² Edward N. Hay. Creating factor comparison key scales by the Per Cent Method. *J. appl. Psychol.*, 1948, 32, 456-464.

³ See H. E. Garrett. *Great experiments in psychology*. New York: Century Co., 1930, pp. 268-274; and Edward N. Hay. Characteristics of factor comparison job evaluation. *Personnel*, 1946, 22, 370-375.

Table 1
Scale for Evaluating Skill Factor

Job Title	Skill Value in Points (15% intervals)
Supervisor of Construction	93
Credit Analyst	81
Supervisor Trust Accounting	70
Head Bookkeeper	61
Cost Accounting Clerk	53
Real Estate Maintenance Inspector	47
Senior Collection Clerk	41
Junior Branch Teller	36
Transcribing Machine Operator	31
Telephone Operator	27
Stock Clerk	23
File Clerk	20
Note Counter	17
Elevator Operator	15
Counter Girl	13
Cleaner	11
Bus Boy	10

then compared for their skill requirement with the jobs on the scale shown in Table 1. For example, the job of Multigraph Operator is studied for the skill requirement as outlined in the factor definitions. Examination indicates that the level of skill required for this job is most nearly the same as for Telephone Operator, and that it requires more skill than Stock Clerk and less than Transcribing Machine Operator. This three-level comparison places the skill requirement at the level of Telephone Operator, which gives it 27 points. In the same manner the requirement of the new job of Multigraph Operator is compared with the jobs on the other factor scales and is given points accordingly. The sum total of the point values for all factors is the point value

for the job as a whole. By entering a specially constructed table of point and salary values we find the salary grade appropriate for the new job of Multigraph Operator.

In factor comparison job evaluation, differences in job difficulty at the threshold of discrimination increase by uniformly equal increments or steps while their salaries increase by a uniform ratio. In working with job evaluation data the writer noticed that Weber's Law seemed to apply and decided to experiment to see if the value of the constant in

Weber's Law, $\frac{dR}{R} = C$, could be determined.

As expressed by Wundt, "all judgments are governed by the general principle of relativity: changes are estimated in terms of the thing which has been changed." Weber put it this way: "in comparing objects we perceive not the actual difference between them but the ratio of this difference to the magnitudes of the two objects compared." That is to say, the observed difference between two objects is not absolute and independent of the objects themselves, but is relative to their size and is a constant fraction of one of them.

This constant fraction, or "difference limen," must be discovered by experiment, and is usually the magnitude at which 75% of the judgments agree. Weber found that for weights lifted by hand it was 1/30 and for the length of lines 1/100. In factor comparison job evaluation it is about 1/7 or 15 per cent. Data presented here show that "just noticeable differences" which were correctly discriminated by trained job raters about three-fourths of the time were 15 per cent apart in salary value of the factor intervals.

In creating the five factor scales in one company, by apportioning the values of the

Table 2
Agreement Between Job Evaluation Estimates by All Evaluators of a Committee and the Final Values they Adopted

Co.	Industry	Number of:				% Agreement
		Jobs	Factors	Evaluators	Judgments	
1 (1947)	Bank	91	5	9	4095	75
1 (1946)	Bank	89	5	9	4005	73
2 (1949)	Life Ins.	300	3	6	5400	73
2 (1947)	Life Ins.	106	3	8	2544	72

Table 3

Agreement Between Job Evaluation Estimates by All Evaluators of a Committee and the Final Values they Adopted

Co.	Industry	Number of:				% Agreement
		Jobs	Factors	Evaluators	Judgments	
1	Bank	112	5	10	5,600	66
6	Utility	382	5	8	15,280	62
8	Metal Mfg.	356	5	8	14,240	63
10	Life Ins.	140	5	7	4,900	61
11	Textile	60	5	5	1,500	55

"key jobs" to these five factors, the intervals on the scale were adjusted to uniform percentage increments, in accordance with the evidence that judgments formed by the use of these scales followed Weber's Law. At that time it was thought that the different factors required different rates of progression. For two factors the progression rate was set at 15 per cent, for another at $11\frac{1}{2}$ per cent, for another at 20 per cent, and for the fifth at 50 per cent. After working with these scales for a number of years it became evident that the first two, using 15 per cent progression, were more satisfactory than the others. The last two scales mentioned appeared to have too coarse intervals, and these scales were re-designed at 15 per cent intervals.

Table 2 shows the results for committees of job evaluators in different companies, all of whom had had more than one year's experience using factor comparison job rating. Each member of the committee works independently and submits his ratings in writing. The next day a meeting of all evaluators is held and each member is given a summary sheet showing the evaluations submitted by all participants. Differences are discussed and agreement is reached on a satisfactory set of ratings. The number of agreements and of differences between the ratings by each person and the final ratings that are adopted by the committee are tallied, which gives percentages of agreement for each member. Table 2 shows the average percentages of agreement for the members of the committees. In the company referred to in the first line of Table 2, for example, 91 jobs were rated in relation to each of 5 factors by each of the 9 members of the evaluation committee. This process involved 4095 separate

judgments. In 75% of these judgments the rating of the individual evaluators agreed with the final value adopted by the committee. Table 2 shows that experienced judges perceived the same differences in job factors about 75 per cent of the time. Since these scales were constructed with 15 per cent increments, these data tend to confirm the finding that 15 per cent was about the right difference in value of the factors.

Table 3 shows the data from five companies in different industries collected during their training periods. These data represent the judgments of relatively inexperienced committees. The percentages of agreement are smaller, as might be expected, ranging from 55 to 66 per cent as against 72 to 75 per cent in the cases of the experienced committees shown in Table 2.

Summary

In factor comparison job evaluation, jobs are "rated" for the assignment of salary or wage rates by comparing them, one at a time, with other jobs with respect to each of several factors or elements common to all jobs. These factor values constitute "scales," each of which consists of an ascending series of values representing the relative salary value of that factor for a job. This ascending series of values progresses by intervals of approximately 15 per cent. The process of comparing one job with another by factors seems to follow Weber's Law, in which the observed differences are a constant percentage of one of the job factors.

*Received November 23, 1949.
Early publication.*

Construction and Use of Verbal Analogy Items

Abraham S. Levine

Human Resources Research Center, Lackland Air Force Base, San Antonio, Texas

World War I saw the emergence of the verbal analogy item in its present form, when it was included as a subtest in the Army Alpha. Yerkes (13) credited Thurstone, Otis and Bingham for the development of this type of item. Since then, the verbal analogy item has been incorporated into a host of widely used tests including the ACE Psychological Examination, National Intelligence Test, Ohio State Psychological Test and the 1937 Revised Stanford-Binet Test. Although it has been commonly used as a subtest, the verbal analogy item constitutes the sole type of item in the Miller Analogies Test, Form G, which is now being widely used as a selection instrument for graduate students.

The analogy form of question is usually of the recognition type consisting of four or five alternative answers, only one of which is correct, and the examinee is required to choose the correct answer. The following item is representative of this form:

Circle is to square as sphere is to (circumference, cube, round, corners, ball).

Although few test makers who employ analogies have been concerned with the theoretical basis for this item, justification for it may be found in the last two of Spearman's three basic laws of cognition dealing with the eduction of relations and the use of the educed relation to discover the correlate. Spearman (9, p. 180) describes the process as follows: "First a pair of ideas is given between which a relation has to be cognized; and then this relation has to be applied to a third idea, so as to generate a fourth one called the correlate." This is in effect what is demanded by a good analogy item.

Several kinds of content have been introduced into the analogies form. Thus, in addition to verbal analogies, there are tests made up of spatial or figure analogies and arithmetic analogies. Perhaps the most unique departure from the traditional type of analogy item was put forth in a test discussed by El Koussy (3).

He suggested that various data indicated that *g* may best be measured by tests of primary perception. In order to eliminate the effect of spatial as well as verbal and number factors he advocated the use of a new test, Greys Analogy Test, which employs varying intensities of grey in analogy form.

However intriguing these other analogy ventures may be, the scope of this paper has been limited to verbal analogies since the available published data on the other forms is as yet too meager to permit the adequate formulation of rules for construction and use.

The general problem of validity of verbal analogy tests may be broken down into three aspects in terms of the research literature on the problem. First, the question may be asked as to how well analogies intercorrelate with other verbal tests, such as vocabulary and opposites, in order to obtain some estimate regarding the saturation of verbal analogies with verbal factors that are measured by other kinds of items. Perhaps the most important aspect of the validity problem is the question of intercorrelation of analogies with tests of general ability. In other words, will a test of verbal analogies by itself give you a good estimate of general academic ability? The factorial composition of verbal analogy tests may be regarded as still another aspect of the validity problem.

With regard to the problem of intercorrelations of verbal analogies with other verbal tests, there is some apparently contradictory evidence. Kelley (6) cites an early research by Agnes Rogers using as subjects 61 entering high school girls with a mean age of 14.6 and an age range of 12-10 to 16-11. Rogers' data showed that the correlation between analogies and logical opposites was only $+.26$. Other low or moderate intercorrelations were obtained by Weisenburg, Roe and McBride (11) employing 70 patients from orthopedic and surgical wards of three hospitals in Philadelphia carefully selected so as to be a fairly representa-

tive sample of the city's population. All subjects used English as their native tongue; ages ranged from 10 to 59. A correlation of $+ .52$ was obtained between printed verbal analogy tests and the Stanford-Binet Vocabulary subtest, and one of $+ .31$ was reported with Thorndike word knowledge. Much higher correlations are reported by both Schneck (8) and Anastasi (1) who both used analogies tests on college students. Schneck reports a $+ .68$ with Vocabulary and a $+ .71$ with Opposites. Anastasi obtained a $+ .65$ with Vocabulary. A plausible interpretation aimed at reconciliation of these apparently contradictory results would be to regard verbal analogy tests as measuring primarily a general ability akin to Spearmanian g at the lower ages and I.Q. levels, and verbal ability at the higher levels. Thus it is possible that difficulty is obtained at the lower levels largely by relational complexity, and at the upper levels most of the variance is attributable to vocabulary complexity. This point bears investigation and it does not gainsay the possibility of adding complexity even on the college and graduate student levels by the addition of relational complexity rather than by abstruse vocabulary as has been more often the case. Indeed, Heim (5) and Hebb (4) independently have attempted to develop analogy tests whose upper sensitivity would be due to something other than vocabulary difficulty.

To what extent do verbal analogies measure what tests of general ability measure? Terman and Merrill (10) included analogy items in the 1937 Revised Stanford-Binet, declaring it to be one of the most satisfactory tests since it correlated uniformly high (average $+ .82$) at each age level with the composite score. In the previously mentioned Weisenburg, Roe and McBride study (11) the following correlations were obtained with the printed analogies test: Stanford-Binet $+ .72$, Completion Beta $+ .63$, Pintner Nonlanguage $+ .55$, Porteus Maze $+ .62$. Such evidence indicates that verbal analogies, at least for a heterogeneous population, measures largely what more general, even non-verbal, tests measure.

Whatever evidence there is available on the problem of factorial composition of verbal analogies points to at least one verbal factor and often another factor such as reasoning or deduc-

tion. The results of these factor analyses are not in complete agreement probably due to differences in the tests and populations used as well as methods of analysis and interpretation. However, for the most part, the factorial studies tend to agree with the logical analysis into verbal and reasoning components.

The reliability coefficients reported for analogy tests tend to run somewhat lower than for other verbal tests. They characteristically bounce around in the .70's and .80's as compared with the .80's and .90's obtained for such tests as vocabulary, opposites, similarities, completion, etc. A combination of some or all of the following factors may tend to explain this finding as well as point to rules for the appropriate use of the analogy item:

1. Greater improvement through practice is reported for analogy tests than for other verbal tests (2).

2. A greater number of zero scores are reported for analogy subtests than for other subtests, particularly for younger and older subjects and subjects with low educational attainment, indicating perhaps greater inherent difficulty in *understanding* the analogy question (2, 13).

3. Analogy subtests tend to be either somewhat shorter or the time allowance is such that the average examinee tends to complete fewer items.

4. The analogy form of question is perhaps subject to greater ambiguity and equivocality of interpretation than other types of items. An important point stressed by Paterson (7), and apparently ignored by some analogy item makers, is that the first two terms should always bear the same relationship to each other as the third and fourth term bear. Mixed and reversed analogies contribute to the ambiguity of the item and convert what would otherwise be an exercise in relational thinking to a sort of guessing game.

The discrimination of a verbal analogy item may be improved by increasing the degree of association between the wrong alternatives and the third term in the item. Zirkle (12) reports a study where the distractors in the analogy items were subjectively divided into three groups on the basis of degree of closeness of association with the third word regardless of the first two words, and analysis showed that

as the distractors went from little association to greatest association there was an increased tendency to choose them. This tendency was greatest for persons scoring in the lowest 27% and less for persons scoring in the highest 27%.

On the basis of the foregoing summary of empirical findings and logical considerations, the following tentative rules may be formulated for the construction and use of verbal analogy items:

1. Sufficient fore-exercises should be included to insure adequate comprehension of the analogy form of question. Such a precautionary measure will tend to reduce sharply the number of zero scores and minimize the amount of practice effect.

2. Items should be arranged in order of difficulty with the first few items constituting graduated practice in analogy thinking as well as serving as shock absorbers.

3. Avoid mixed and reversed analogies by making sure that the first two terms are related to each other in the same fashion as are the last two terms.

4. Distractors should be made as plausible as possible, and wherever feasible, at least one of the distractors should be closely associated with the third term in meaning.

5. The test should include enough items and the time limit should be long enough to enable examinees to answer sufficient items to insure adequate reliability. It is probably desirable to use verbal analogy items in a power test rather than a speed test since the variable of set may play a more important role in understanding what is required than in other types of items.

6. Since it is more difficult to foresee possible interpretations of analogy questions, a rigid selection on the basis of item analysis data is perhaps even more important for analogies than for other types of items.

7. It is perhaps more desirable in measuring general ability to attain difficulty by means of complexity of relationship rather than through

the use of uncommon vocabulary. It should be pointed out, however, that the superiority of the former type of item as a predictor of academic success has not as yet been demonstrated.

8. For optimal effectiveness, the use of the analogy item may best be restricted to a college or professional school level of young adults. Younger and older subjects, as well as less intelligent and poorly educated examinees, tend to experience the most difficulty in understanding the analogy form of question and therefore it should be used with great caution for these segments of the general population.

Received June 10, 1949.

References

1. Anastasi, Anne. Further studies on memory factor. *Arch. Psychol.*, 1932, No. 142, 1-60.
2. De Weerd, E. H. The transfer effect of practice in related functions upon a group intelligence test. *Sch. & Soc.*, 1927, 25, 438-440.
3. El Koussy, A. H. A note on the Greys Analogy Test. *Brit. J. educ. Psychol.*, 1934, 4, 294-295.
4. Hebb, D. O. Verbal test material independent of special vocabulary difficulty. *J. educ. Psychol.*, 1942, 33, 691-696.
5. Heim, A. W. An attempt to test high grade intelligence. *Brit. J. Psychol.*, 1947, 37, 70-81.
6. Kelley, T. L. *Crossroads in the mind of man*. Stanford U., California: Stanford U. Press, 1928.
7. Paterson, D. G. *Preparation and use of new-type examinations*. New York: World Book Co., 1926.
8. Schneek, M. R. The measurement of verbal and numerical abilities. *Arch. Psychol.*, 1929, No. 107, 1-49.
9. Spearman, C. *Abilities of man*. New York: Macmillan Co., 1927.
10. Terman, L. M., and Merrill, M. A. *Measuring intelligence*. New York: Houghton Mifflin Co., 1937.
11. Weisenburg, T., Roe, Anne, and McBride, Katharine E. *Adult intelligence*. New York: Commonwealth Fund, 1936.
12. Zirkle, G. A. An analytic study of the multiple choice analogies test item. *J. educ. Psychol.*, 1946, 37, 427-435.
13. *Memoirs of the National Academy of Sciences*, V. 15. Washington: Govt. Printing Office, 1921.

Spread of Vocational Interests and General Adjustment Status *

Samuel F. Klugman

Veterans Administration, Philadelphia, Pa.

Opinions as to the relationship between personality inventory scores and vocational interest scores vary from one extreme to the other. On one hand, for example, there are these statements by Tyler and Fowler, respectively. The former states: "Scores indicating neurotic tendencies show no appreciable relationship to any kind of interest scores" (20, p. 67). The latter writes: "General agreement seems to exist, too, that interest test scores are not a dependable basis for conclusions about a student's attitudes and adjustment" (9, p. 26). On the other hand, Darley claims, "Men with primary interest patterns in technical occupations . . . seem to have markedly poorer social adjustments . . . (and) they show a tendency for consistently better home, health, and emotional adjustments" (7, p. 470-472). In a similar manner he differentiates individuals with Verbal and Social Service A patterns from those with Business Contact and Social Service B patterns. At a later date, based on 80 maladjusted adult cases, he maintains that the data fit "into the hypothesis of interests as an outgrowth of personality" (8, p. 58). Sarbin and Berdie found sufficient relationship to say, "It is possible in the absence of other interest measurement to use the Allport-Vernon Scale (a personality inventory) to approximate certain occupational interest types as measured by the Strong test" (15, p. 296). Between these two extremes, is this statement by Alteneder: "Relationship between vocational interests and intelligence and personality traits indicate some interesting trends" (2, p. 459). Tussing also declares, "From this study it does appear that a prediction of self-confidence and of

sociability can be made with a fair amount of accuracy . . ." (19, p. 72). However, he found no relationship between Strong's Scales and Bell's Home, Health, and Emotional Adjustment Areas. Strong, too, feels that "interests as measured today are related somewhat to certain personality factors and attitudes" (16, p. 341).

It is fairly obvious that there is a need for further investigation of this problem. Furthermore, the test which has been used most frequently is the Strong Vocational Interest Blank. Perhaps the use of another test would yield more consistent results, and toward this end the Kuder Preference Record was employed in this study.

Statement of the Problem

The chief purpose of this study was to determine whether any relationship existed among veterans, between "general adjustment status" as measured by the total score of the Bell Adjustment Inventory (3, p. 1) and "spread of vocational interests" based on Kuder Preference Record profile scores (13).

So far as this writer is aware, "spread of interest" as used in this investigation is a new concept and is not to be confused with Berdie's "range of interests" which is determined merely by counting the number of interests claimed from a list of 22 recreational items (4, p. 269). Essentially, "spread of interest" is similar to Wechsler's technique for determining the significance of a varying score by comparing it with the individual's sub-test mean (21, pp. 148-149). In this instance, however, it refers to the sum of the deviations of all scores from the individual's mean based on the assignment of standard scores (as per Bingham, Chapter XIX) to the nine percentile scores of the Kuder Preference Record. The greater the spread score the more definite are one's likes and dislikes. Stated in another way, one who has a profile with only very high and very low scores is a person with a more

* Published with the permission of the Chief Medical Director, Department of Medicine and Surgery, Veterans Administration, who assumes no responsibility for the opinions nor conclusions by the author. The writer wishes to acknowledge his indebtedness to Dr. Hans C. Gordon, Division of Educational Research of the Philadelphia School System, for reading this manuscript and offering several constructive suggestions. He assumes, of course, no responsibility for any of the writer's statements.

definite pattern of interests than one whose scores fall closer to the mean.

Besides the null hypothesis, there are at least two others which may be stated:

1. The more definite one's likes and dislikes are, resulting in a greater spread of one's vocational interest pattern (i.e., the greater the sum of the deviations from his mean), the better adjusted is he.

2. The more one's interests are distributed in a horizontal manner over a wide area, embracing many fields (not liking or disliking many fields with too much intensity, resulting in smaller deviations), the better adjusted is he.

In addition to the purpose stated above, this study seeks to determine whether such factors as age, education, and intelligence, as well as spread of vocational interests and general adjustment status, are related to the various occupational interest areas measured by the Kuder Preference Record.

Procedure

During the year 1947, by random selection, the records of 108 veterans who had applied, or had been sent by the Regional Office, for advisement to the Veterans Administration-University of Pennsylvania Guidance Center, where the present writer was serving as Chief, were used for this study. These records contained at least

the following pertinent information: age, sex, formal schooling, race, intelligence quotient (as obtained on the California Mental Maturity test, advanced form, timed), vocational interest profile (as per Kuder Preference Record) and personality inventory scores (Bell Adjustment Inventory, adult form).

The Subjects

The veterans employed in this study were white, native-born males who had a mean age of 24.5, S.D. 3.99; had completed 11.3, S.D. 1.37, grades of schooling; and had an average IQ rating of 111.9, S.D. 14.78. This higher than average mental ability was due to two factors: (1) in order to minimize possible misunderstanding of the language elements making up the tests only those who had completed at least the Eighth Grade were selected; and (2) many university students dropped in for advisement purposes because the guidance center was so conveniently located for them.

Results

The obtained data for the interest test were not essentially different from the norm population. Furthermore, these data were generally in agreement with Rose's veterans who were also tested at a university Veterans Administration Guidance Center (14).

Table 1
Scattergram Showing Relationships Between Kuder Spread-of-Interest Scores and Bell General Adjustment Status Scores

Bell General Adjustment Status Scores*												
	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	101-110	Total
Kuder Spread-of-Interest Scores**	7.01-7.51		1									1
	6.51-7.00	1			1							2
	6.01-6.50	1		1	1	1						4
	5.51-6.00	1		2	1		2		1	2	1	10
	5.01-5.50	1	3	3	1	4	1	1				14
	4.51-5.00	1	2	5	1	4	1	2				16
	4.01-4.50	1	1	5	2	3			1			13
	3.51-4.00	1	3	3			3	1	1			12
	3.01-3.50			3	2	2	3	1				11
	2.51-3.00				1	4	1					6
	2.01-2.50	1	1	3	2	2	3	1				13
	1.51-2.00	2	1	1				1		1		6
	Total	10	12	26	11	20	13	9	2	2	2	1

* Based on total raw scores.

** Based on standard score deviations from individual means.

Table 1 is a scattergram which shows how the various scores for General Adjustment Status and for spread of vocational interests are distributed. By inspection, it may be easily noted that the scores are widely scattered and do not follow a consistent pattern. This is further verified by the calculation of a product-moment correlation of $.01 \pm .07$. It may be concluded, then, that no relationship was found between spread-of-interest scores (as measured by our suggested technique) and General Adjustment scores (as measured by the Bell Adjustment Inventory).

In addition to the above product-moment correlation, others were determined for spread-of-interest scores and the three factors of education (i.e. final grade completed), age, and IQ rating. For education, a correlation of .31 was obtained; for age .29; and for IQ rating .27. In all instances, the P.E. was .06. These correlations are significant at the .01 level and each relationship may be described as "present but slight." It may be said, then, that the veterans who were older, more intelligent, and possessed more formal schooling tended to have greater spread-of-interest scores than those who were younger, less intelligent, and had completed fewer years of schooling.

Table 2 gives a series of relationships between five factors and the various Kuder interest areas. In dividing the group into halves, those with percentile scores above 50 for a particular interest area were placed in the Stronger interest group and those with percentiles below 50 in the Weaker interest group. Those with scores of exactly 50 were put in the half with the lesser number of cases in order to keep the two groups as nearly alike in size as possible. The interpretation of critical ratios generally follows Guilford (10, p. 61). For those interested in interpretations based on Student's distribution for a group as large as this, a C.R. of 1.98 indicates confidence at the 5 per cent level; 2.36 at the 2 per cent level; and 2.63 at the 1 per cent level. In terms of factors studied, these results were obtained:

1. In no area did spread of vocational interest scores yield differences beyond reasonable doubt between those with stronger and weaker vocational interests.
2. Reliably better adjustment scores were

obtained by those with stronger Scientific interest scores. Those with weaker Artistic interest scores earned scores which indicated fair certainty of difference in their favor. Had the Social Service interest scores yielded a C.R. of 1.98 instead of 1.92, it would have been significant at the 5 per cent level of confidence in favor of those in the weaker half. It is interesting to note, in passing, that Kimber had found that Bible students had "less freedom from nervous symptoms than the average person, the prevailing interest was in social service, and there was a noticeable lack of interest in computational and clerical activities" (11, p. 233). It will be seen that the present findings are in agreement because those with stronger Social Service, weaker Computational, and weaker Clerical interests had poorer adjustment scores. However, none of the differences is significant although, as pointed out above, there is a relatively high level of confidence for the Social Service area.

3. Only in the Mechanical area was there obtained a difference large enough to point to a relationship with IQ ratings. With fair certainty, it was found that those with stronger Mechanical interests earned lower IQ ratings. An interesting result may be noted with reference to the Scientific category; namely, IQ's differentiated least between those who liked and those who disliked scientific activities as stated in the Kuder test. This is contrary to a conclusion drawn by Super (17, p. 23) from Strong's findings (16). On the other hand, it supports his statement concerning Social Service interests (p. 24).

These findings are essentially in agreement with those of Adkins and Kuder who, using a test of intelligence, stated that "... the interpretation of preference scores as indicative of the presence or absence of special abilities is unwarranted by the results of this investigation" (1, p. 261). Similarly, Strong writes that "the differences are too slight to be usable in prediction of individual cases" (16, p. 334). Triggs' highest coefficient of correlation between Kuder interest scores and ACE total scores was only .27 (Literary category), once again indicating low relationship (18, p. 349).

4. Age apparently bore no relationship to any of the interest areas, since all critical ratios were even below the lower criterion of 2.00.

Table 2
Differences Between Stronger and Weaker Interest Groups for Various Factors

	N	Spread of Interests			General Adjustment			I.Q. Ratings			Age			Education		
		Mean	S.D.	C.R.	Mean	S.D.	C.R.	Mean	S.D.	C.R.	Mean	S.D.	C.R.	Mean	S.D.	C.R.
Mechanical	S*	3.6	1.14	1.88	40.4	20.59	1.05	108.9	13.42	2.16	25.5	4.24	1.76	10.9	1.61	2.47
	W*	4.7	1.25		36.1	21.12		114.9	15.65		23.6	6.78		11.6	1.41	
Computational	S	4.3	1.06	.51	34.6	17.75	1.61	114.6	16.49	1.79	24.3	5.48	.45	11.7	1.55	2.51
	W	4.0	1.49		41.8	24.31		109.2	14.73		24.7	3.99		10.9	1.64	
Scientific	S	4.3	1.28	.42	31.7	20.35	3.80	112.1	13.23	.11	24.6	4.45	.97	11.3	1.56	.19
	W	4.0	1.29		45.5	17.52		111.7	14.87		24.5	4.90		11.2	1.62	
Persuasive	S	4.1	1.26	.37	36.5	22.11	.96	113.5	14.25	1.30	24.0	3.63	1.72	11.4	1.30	.97
	W	4.3	1.33		40.4	20.82		109.8	15.07		25.2	3.41		11.1	1.94	
Artistic	S	4.0	1.33	.64	42.2	23.02	1.99	110.1	15.03	1.27	24.9	2.58	1.00	11.1	1.80	1.26
	W	4.4	1.25		34.1	18.97		113.8	15.68		24.2	4.37		11.5	1.48	
Literary	S	4.2	1.32	.25	36.1	24.57	.94	114.1	19.73	1.44	23.9	3.38	1.51	11.5	1.22	1.37
	W	4.1	1.25		40.2	20.95		109.3	14.86		25.1	4.69		11.1	1.85	
Musical	S	4.3	1.42	.32	40.8	22.18	1.26	112.1	15.65	.11	24.4	4.04	.44	11.1	1.62	1.31
	W	4.1	1.18		35.6	20.51		111.7	19.33		24.7	4.17		11.5	1.44	
Social Service	S	4.2	1.38	.29	41.5	23.46	1.92	112.4	18.65	.44	24.6	4.82	.19	11.2	1.70	.19
	W	4.1	1.21		33.8	18.18		111.1	12.22		24.4	3.77		11.3	1.52	
Clerical	S	4.4	1.28	.70	36.9	19.29	.67	109.5	15.36	1.53	24.9	2.16	1.39	10.9	1.50	2.37
	W	3.9	1.27		39.6	22.61		114.2	16.31		24.6	4.20		11.7	1.67	

* S: Stronger interest group (above 50th percentile).

* W: Weaker interest group (below 50th percentile).

5. Education seemed to be significantly related to more interest areas than any other factor investigated in this study. Those with weaker Mechanical, weaker Clerical, and stronger Computational interest scores had significantly completed more years of formal schooling. In this connection, Crosby obtained results that "show a positive relationship between interests in certain scales of the Preference Record and achievement in some school subjects" (6, p. 103). Triggs also found relationships which "tend to indicate that interests and achievement are not totally unrelated" (18, p. 354). The present writer, on an earlier occasion, when he investigated the effect of schooling on clerical interests, stated, ". . . it is not maturation which is responsible for the improvement in scores . . . but probably schooling" (12, p. 95).

The data in Table 3 are based on the findings reported in Table 2. Whereas, in the earlier table the relationships of the various factors are noted *within* each of the interest categories (i.e., those with stronger are compared with those having weaker interests), the present table is concerned with the relationships of these factors *among* the interest categories.

Furthermore, the data in this table deal with positive interest alone in the sense that comparisons were made only between the stronger halves (above 50th percentile) of the interest categories. Finally, in Table 3 are the significant differences only, i.e., those critical ratios that are significant at least at the 5 per cent level of confidence. The sole exception is so close to the criterion of 1.98 that it is not amiss to include it in this table.

One important fact to note in this table is that the Mechanical interest category appears to be significantly related to all factors. Those with stronger Mechanical interests are less definite in their likes and dislikes of other vocational interests (as measured by the suggested technique) than are those with stronger interests in the Clerical, Computational, Scientific, Social Service, Musical, and Literary areas. General Adjustment status appears to be poorer for them (since lower inventory scores are more desirable) than for those with strong Scientific interests. IQ ratings were lower for them, too, than for those with stronger interests in the Computational category. In addition, those with stronger Mechanical interests were reliably older than those

Table 3
Significant Inter-Category Difference Means for the Various Factors

Factor	Interest Categories	Inter-Category Difference*	C.R.
Spread of Interests	Mechanical and Clerical	.75	3.26
	Mechanical and Computational	.69	3.24
	Mechanical and Scientific	.67	2.90
	Mechanical and Social Service	.62	2.57
	Mechanical and Musical	.63	2.54
	Mechanical and Literary	.61	2.52
General Adjustment Status	Scientific and Artistic	10.52	2.56
	Scientific and Mechanical	8.68	2.45
	Scientific and Social Service	9.79	2.36
	Scientific and Musical	9.15	2.26
	Computational and Artistic	7.63	2.10
IQ Ratings	Mechanical and Computational	5.75	1.97
Age	Literary and Mechanical	1.51	2.07
	Persuasive and Mechanical	1.47	1.99
Education	Mechanical and Computational	.75	2.45
	Mechanical and Literary	.55	2.02

* Higher mean found in second interest category of each pair.

with positive Literary and Persuasive interests. They also had less formal schooling than those with stronger Computational and Literary interests.

Those with stronger interests in the Scientific area were better adjusted, to a fair degree of certainty, than those with interests in Artistic, Social Service, Mechanical, and Musical fields. Similarly, those with positive Computational interests were emotionally more stable than those with Artistic ones.

It is interesting to note that Education, which was the factor related to most intra-category differences, did not play so important a role in inter-category differences.

Summary

The records of 108 veterans who had applied for advisement at the Veterans Administration-University of Pennsylvania Guidance Center were selected at random in order to determine what, if any, relationship existed between vocational interest and personality inventory scores as measured by the Kuder Preference Record and the Bell Adjustment Inventory, respectively. Of special interest was the determination of the value of a new measure of spread of vocational interests based on mean deviations of an individual's standard scores. In addition, the factors of age, IQ ratings, and education were studied in relation to interests.

The following findings were obtained:

1. Spread of vocational interest, as measured in this study, showed no relationship with General Adjustment Status. This lack of relationship was also found in a comparison between those with stronger and those with weaker interests in the various areas.
2. There was a slight tendency for the older, more educated, and more intelligent veterans to have greater spread-of-interest scores than the younger, less educated, and less intelligent ones.
3. Spread-of-interest scores succeeded in differentiating between those with stronger Mechanical interests and stronger Clerical, Computational, Scientific, Social Service, Musical, and Literary to a degree of fair certainty at least. In all instances, the mean spread score of stronger Mechanical interests was less than the mean spread score of the others.

4. Veterans with stronger Scientific interest scores earned reliably better adjustment scores, as did those with weaker Artistic interest scores (but to a lesser degree), than their opposite halves. General Adjustment Status for those with stronger Scientific interests was significantly better than for those with stronger Artistic, Social Service, Mechanical, and Musical interests. To a lesser extent, this is also true for those with positive Computational interests over positive Artistic ones.

5. IQ ratings did not differentiate within any of the interest categories except for the Mechanical area where those with weaker scores earned higher IQ's to a degree of fair reliability. A significant difference (at the 5 per cent level of confidence) was obtained, on the basis of these scores, between those with positive Mechanical and Computational interests.

6. The factor of age did not succeed in differentiating any of the veterans on the basis of stronger versus weaker vocational interest scores. Significant age differences (at the 5 per cent level) were found between positive Mechanical and positive Literary and Persuasive interest areas respectively. Those with stronger Mechanical interests were older than those with high interests in the other two fields.

7. Education, more than any of the factors studied, was related to a significant degree when studied on an intra-category basis. However, this factor was not so effective in differentiating among those with positive interests.

8. The Persuasive, Literary, and Musical areas of interest were found not to be related with certainty to any of the factors investigated when studied on an intra-category basis. Attention is called to the Social Service area which is barely outside the 5 per cent level of confidence in relation to General Adjustment Status. When investigated on the basis of positive inter-category interests, however, all were found to be related to one or more of the factors.

9. The Mechanical interest area was related to all the factors on an inter-category basis with at least fair certainty. On an intra-category basis, the same relationship was found for all factors except spread-of-interest scores.

10. Relationship between a personality inventory total score (Bell Adjustment) and

vocational interest test scores does not appear to be greater when the Kuder Preference Record is used than when the Strong Vocational Interest Blank is employed. To use Strong's descriptive statement, the scores are "related somewhat." Therefore, it does not appear to be feasible, at this time, for vocational counselors, clinical psychologists, and others who make use of these types of tests, to infer the presence of definite relationships except in a limited manner.

Received June 20, 1949.

References

- Adkins, D. C., and Kuder, G. F. The relation of primary mental abilities to activity preferences. *Psychometrika*, 1940, 5, 251-262.
- Altener, L. E. The value of intelligence, personality, and vocational interests in a guidance program. *J. ed. Psychol.*, 1940, 31, 449-459.
- Bell, H. *Manual for the Adjustment Inventory*. Stanford University Press, 1935.
- Berdie, R. F. Range of interests. *J. appl. Psychol.*, 1945, 29, 268-281.
- Bingham, W. V. D. *Aptitudes and aptitude testing*. New York: Harpers, 1937.
- Crosby, R. C. Scholastic achievement and measured interests. *J. appl. Psychol.*, 1943, 27, 101-104.
- Darley, J. G. A preliminary study of relations between attitude, adjustment, and vocational interest tests. *J. ed. Psychol.*, 1938, 29, 467-473.
- Darley, J. G. *Clinical aspects and interpretation of the Strong Vocational Interest Blank*. New York: Psychological Corp., 1941.
- Fowler, F. M. Interest measurement,—questions and answers. *Sch. Life*, 1945, 28, 25-29.
- Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1936.
- Kimber, J. A. M. Interests and personality traits of Bible institute students. *J. soc. Psychol.*, 1947, 26, 225-233.
- Klugman, S. F. Test scores for clerical aptitude and interests before and after a year of schooling. *J. genet. Psychol.*, 1944, 65, 89-96.
- Kuder, G. F. *Revised Manual for the Kuder Preference Record*. Chicago: Science Research Associates, 1946.
- Rose, W. A comparison of relative interest in occupational groupings and activity interests as measured by the Kuder Preference Record. *Occupations*, 1948, 26, 302-307.
- Sarbin, T. R., and Berdie, R. F. Relation of measured interest to the Allport-Vernon study of values. *J. appl. Psychol.*, 1940, 24, 287-296.
- Strong, E. K. *Vocational interests of men and women*. Stanford University Press, 1943.
- Super, D. E. Experience, emotion, and vocational choice. *Occupations*, 1948, 27, 23-27.
- Triggs, F. O. A study of the relation of Kuder Preference Record scores to various other measures. *Educ. & Psychol. Meas.*, 1943, 3, 341-354.
- Tussing, L. An investigation of the possibilities of measuring personality traits with the Strong Vocational Interest Blank. *Educ. & Psychol. Meas.*, 1942, 2, 59-74.
- Tyler, L. E. Relationships between Strong Vocational Interest scores and other attitude and personality factors. *J. appl. Psychol.*, 1945, 29, 58-67.
- Wechsler, D. *The measurement of adult intelligence*. Baltimore: Williams & Wilkins, 1944.

Kuder Interest Patterns of Psychologists *

Malcolm L. Baas

Purdue University

The primary purpose of this study was to secure information concerning motivations, interests, or preferences of professional psychologists, and of graduate students in psychology. Significance of this purpose is reflected in such studies as *The Research Project on the Selection of Clinical Psychologists* (6) being conducted at the University of Michigan under the supervision of E. Lowell Kelly and Donald W. Fiske.

Cognizance of such information may introduce considerations in the guidance of students toward or away from psychology; and also considerations in the evaluation of personality patterns among practicing professional psychologists.

Procedure

Groups Used. Subjects from four divisions of the American Psychological Association were selected for this study. These divisions were: Clinical, Industrial, Counseling and Guidance, and Experimental and Theoretical. From each division, sixty fellows were randomly selected and subsequently contacted. Also participating in the study were groups of graduate students in both Industrial and Clinical Psychology at Purdue University. The graduate students were candidates for either the M.S. or the Ph.D. degree at the time of this study.

Administration of the Kuder Preference Record. The subjects were requested to complete the Kuder Preference Record, Form BM, and return it to the author. From the originally selected professional group of 260, returns permitted a statistical analysis of 29 Industrial, 27 Clinical, 26 Counseling and Guidance, and 29 Experimental and Theoretical Psycho-

logists, a total of 111. This is a 46 per cent return. None of the psychologists considered in this final group was a fellow in more than one of the divisions as listed. From the student groups, 21 Clinical and 25 Industrial Psychology Graduate Students were considered in the final study.

Treatment of Data. Mean raw scores¹ on each of the nine Kuder Scales were obtained for each of the Professional groups, and for the graduate student groups. Similarly, percentiles, standard deviations, and standard errors were obtained.

Using the Kuder Adult Norms for Men, percentile ranks for the mean raw scores were found. The mean scores were also interpreted in terms of norms derived from the entire professional psychologist sample.

To indicate significance of mean-differences between the divisions, Student's *t*-ratios were tabulated. These ratios serve to test the null hypothesis concerning the differences in means between professional groups, and between student and professional groups.

All statistical data have been reported in tabular and profile forms, but because of their extensive length they must be deleted from this paper.

Results

For the entire professional sample (Table 1) two scales exhibited strength of interest. The Scientific and Literary scores were above the 75th percentile of Kuder's Adult norms for men.

The four divisions seem fairly consistent in reflecting the scientific and literary interests. For these two scales, only one significant difference in means occurred. The Scientific Scale difference between the means for the Theoretical and the Clinical Divisions was significant at the 2% level of confidence. Inasmuch as the difference favored the Theo-

* This article is based upon the author's thesis, entitled "A Study of Interest Patterns Among Professional Psychologists," submitted to the Faculty of Purdue University in partial fulfillment of the requirements for the degree of Master of Science, February, 1949. The thesis was directed by Professor Franklin J. Shaw.

¹ Because of the extensive amount of statistical data, the tabulations and profile presentations have been deleted from this article. The original thesis is filed in the Purdue University Library.

Table 1
Kuder Mean Percentiles for All Groups

Group	Scales							
	Mech.	Comp.	Sci.	Pers.	Art.	Lit.	Mus.	Soc. Serv.
Psychologists in general	32	67	82	24	59	86	60	60
Clinical	32	55	74	12	87	80	60	75
Industrial	40	75	82	45	40	87	50	45
Counseling	23	64	77	33	30	86	55	90
Theoretical	35	70	89	11	67	87	75	25
Student Clinical	8	22	32	59	46	76	78	96
Student Industrial	26	70	40	76	12	71	60	78
								30
								19
								40
								32
								30
								19
								49

retical Division, this outcome seems consistent with the divergent patterns of activity of the two fields.

Low interest areas for the professional psychologists were reflected in the mean scores for the Mechanical, the Persuasive, and the Clerical Scales. Significant differences in means were found on the Mechanical Scale, the Counseling Division's mean score being significantly less than the mean score for either the Theoretical or the Industrial Divisions. Levels of confidence for the differences were 5% and 1%, respectively. The Industrial and Counseling Divisions had higher mean scores on the Persuasive Scale than the other divisions. Industrial Psychologists evidenced a stronger interest on this scale than either the Clinical or the Theoretical Divisions, with significance at the 1% level of confidence. On the Clerical Scale, the difference between the means for the Industrial and Clinical Divisions

was significant at the 5% level of confidence, favoring the Industrial group.

The middle-range mean scores also showed significant differences between certain of the tested groups. The Artistic Scale differentiated between the Counseling and Industrial divisions, favoring the Industrial Division at the 5% level of confidence. The same scale differentiated between the Theoretical Division and the Counseling Division. For both these mean-difference comparisons, the Theoretical Division mean score was significantly higher at the 2% level of confidence. No significant differences occurred for the Computational Scale mean-score comparisons, but the Social Service Scale exhibited significant mean differences in all the inter-division comparisons. For the Social Service Scale, mean-differences at the 1% level of confidence favor the Clinical Division over both Industrial and Theoretical Divisions; and the Counseling Division over

Table 2
Profile Statistics for 111 Professional Psychologists

Scale	Mean Raw Score	%-ile M (Kuder Norms)	%-ile M (Sample Norms)	Standard Deviation	S.E. of Mean
Mechanical	68.9	32	50	17.5	1.7
Computational	40.0	67	50	10.1	1.0
Scientific	77.8	82	50	12.7	1.2
Persuasive	57.7	24	50	17.6	1.7
Artistic	48.2	59	50	16.2	1.5
Literary	64.8	86	50	12.1	1.2
Musical	17.4	60	50	8.8	0.8
Social Service	78.2	60	50	22.0	2.0
Clerical	44.7	30	50	12.6	1.1

both Industrial and Theoretical Divisions. The Social Service mean-difference between the Industrial and Theoretical Divisions favors the Industrial Division at the 2% level of confidence. And the Counseling Division's Social Service mean score is significantly higher than the Clinical Division's mean score at the 5% level of confidence.

Significant differences also prevailed between the mean scores of the professional psychologists and their student counterparts. Only the Literary and the Clerical Scales failed to differentiate between the Professional Clinical Psychologists and the Clinical Psychology Graduate Students. With regard to the Professional Industrial Psychologists and the students of Industrial Psychology, four scales failed to show differences in means. The four scales were: The Mechanical, The Computational, The Musical and The Clerical Scales. In view of these results it seems possible that preference patterns are considerably affected by professional experience, although Strong's work (9) would not lend support to this view. It is possible that interest patterns will become more stable and that strong interest areas will become better established as experience contributes to an individual's preferences. Several investigators (1, 2, 4, 7, 11) have indicated, nevertheless, that the *Kuder Preference Record* can be of help in selecting and guiding students.

Summary

In this investigation an effort was made to identify characteristic interest patterns for four different groups of professional psychologists and for two groups of student psychologists. Employing the *Kuder Preference Record*, preferences of groups selected from four American Psychological Association divisions and two groups of psychology graduate students were obtained.

Having received the completed *Kuder Preference Record* forms from sample groups of fellows in the Clinical, the Industrial, the Counseling and Guidance, and the Theoretical and Experimental Divisions, a statistical analysis of these results was made. A similar analysis was performed on the graduate student data.

The findings revealed high scores for all professional psychologists on the Scientific and Literary Scales of the *Kuder Preference Record*. Weak interests were evident on the Mechanical, the Persuasive, and the Clerical Scales. Several significant differences between the groups studied were found on the various scales.

The findings suggest that professional experience may contribute to preference scores, although Strong's work on age changes in interest patterns would not support this. However, this particular investigation has not incorporated an examination of subjects' early interest backgrounds. Age differences, experience differences, and environmental influences should be considered. And the small samples, as well as having used only Purdue University graduate students, are also limiting factors.

Received July 5, 1949.

References

1. Adkins, Dorothy C., and Kuder, G. Frederick. The relation of primary mental abilities to active preferences. *Psychometrika*, 1940, 5, 251-262.
2. Bolanovitch, Daniel J., and Goodman, Charles H. A study of the Kuder Preference Record. *Educ. & Psychol. Meas.*, 1944, 4, 315-326.
3. Crosby, R. C., and Winsor, A. L. The validity of student estimates of their interests. *J. appl. Psychol.*, 1941, 25, 408-414.
4. Hahn, Milton E. Notes on the Kuder Preference Record. *Occupations*, 1945, 33, 467-470.
5. Hahn, Milton E., and Williams, Cornelia T. The measured interests of Marine Corps Women Reservists. *J. appl. Psychol.*, 1945, 29, 198-211.
6. Kelly, E. Lowell, and Fiske, Donald W. *The selection of clinical psychologists*. Ann Arbor, Mich.: University of Michigan, 1948.
7. Kimber, J. A. Morris. Interests and personality traits of Bible Institute students. *J. soc. Psychol.*, 1947, 26, 225-233.
8. Kuder, G. Frederick. *The Kuder Preference Record: A revised manual*. Chicago, Illinois: Science Research Associates, 1946. 30 pages.
9. Strong, E. K., Jr. *Change of interest with age*. Stanford University, California: Stanford University Press, 1931.
10. Strong, E. K., Jr. *Vocational interests of men and women*. Stanford University, California: Stanford University Press, 1943.
11. Yum, K. S. Student preferences in divisional studies and their preferential activities. *J. Psychol.*, 1942, 13, 193-200.

A Study of Shooting Glasses by Means of Firing Accuracy *

Sherman Ross

Bucknell University, Lewisburg, Pennsylvania

The use of "shooting glasses" is apparently a very old idea in the field of hunting or of range firing. In the museum at Fort Necessity, a pair of old-fashioned spectacles with red glass, labelled: "Hunting Glasses, about 1830" are to be found.

There have appeared on the commercial market several different types of glasses, with an assortment of filters, which have been declared to be shooting glasses. The major purpose of these glasses is to provide better resolution of the target, under conditions of glare, haze, etc. The glasses have been produced commercially under such trade names as "Rayban," "Calobar," and "Eaglesight."

The use of filters in "cutting haze" and in providing better resolution of targets is a well-established notion in the field of vision. In a recent report, Verplanck¹ has studied the effectiveness of selected neutral and colored filters in extending the visual range at which neutral targets, viewed against a sky background, may be discriminated in the presence of haze. The results of the experiments indicate that the use of colored or neutral filters did not increase the range at which neutral targets, viewed against a sky background, may be discriminated. The present experiment deals not with the problem of discrimination, but with the complex of visual-psycho-motor variables represented by the actual firing of a rifle at a target. The purpose of the present study was to determine the effectiveness of

several types of plastic filters when used as shooting glasses by a group of skilled Marine Corps riflemen during range firing.

Method and Procedure

It will be convenient to discuss the method and procedure under the following headings: (1) Subjects, (2) Filters, (3) Firing Schedule, and (4) Method of Scoring.

1. *Subjects:* The subjects for the experiment were a group of 21 riflemen with the highest qualifications, selected from a group of 28 range coaches who had reported for duty during the previous week.

2. *Filters:* Six types of filters and one clear control were used. The characteristics of these filters are shown in Table 1. This table shows the Polaroid Corporation code numbers, the polarization properties of the filter, the color, and the transmission characteristics. The plastic filters were placed in identical, rubber mounted frames (Polaroid 1065), and were marked with identifying letters.

3. *Firing Schedule:* The tests ran for three consecutive days. Firing was begun between 0815 and 0845 and ended between 1100 and 1130. The weather was clear and sunny, and there were relatively few clouds during the tests. The sun came from over the right shoulders of the men on the firing line. The subjects were divided into three groups of seven men each. This was done for convenience in handling the goggles and in scoring the targets. Sighting in of the rifles was done each day previous to firing. Eight shots, slow fire, prone at 200 yards with a standard target, constituted a string. The M-1 rifle was used throughout the tests. The daily plan of the firing for each group of seven men is presented in Table 2. As will be seen, the firing was done in alternate strings: with goggles and without goggles. However, when goggles were used, each man used a different type. After Relay 1 was fired by the first group of seven men, the

* This study was performed at the Medical Field Research Laboratory, Camp Lejeune, North Carolina, during the fall of 1944, with the cooperation of the Range Battalion, Camp Lejeune. Opinions expressed in this report are those of the author and are not to be construed as official or as representing the views of the Navy Department or of the Naval Service at large. The writer should like to express his thanks to S. B. Williams, College of William and Mary, and to S. B. Lyerly, University of North Carolina, for their help on the manuscript.

¹ Verplanck, W. S. A field test of the use of filters in penetrating haze. Report on Bu M & S Research Project HM-011-003, dated 6 June 1947. From the Medical Research Department, U. S. Submarine Base, New London, Conn. Pp. 11.

Table 1
Showing the Characteristics of the Filters

Filter	Polaroid Corp. No.	Polarization	Color	Transmission Characteristics
A	XC92	Non-Polarizing	Neutral	Essentially complete transmission throughout visible region except for reflection losses
B	XN10	Non-Polarizing	Neutral	Approximately 10% overall visual transmission
C	HN32	Polarizing	Neutral	Approximately 32% overall visual transmission
D	MG30	Polarizing	Neutral	Approximately 30% overall visual transmission
E	XY68 ¹	Non-Polarizing	Amber	Approximately 68% overall visual transmission
F	XG30 ²	Non-Polarizing	Green	Approximately 30% overall visual transmission
G	XO50 ³	Non-Polarizing	Light Red	Approximately 50% overall visual transmission

¹ This filter selected because of its resemblance to the commercially produced "Rifleite" shooting glass.

² This filter selected due to its resemblance to the commercially produced "Rayban" and "Calobar" glasses.

³ This filter selected since it resembled the commercially produced "Eaglesight" shooting glass.

second group of seven men fired Relay 1. After Relay 1 was fired by the third group of seven subjects, the first group fired Relay 2, etc. The firing for a single day for each subject therefore was composed of 14 strings of eight shots, seven strings without goggles, and one string with each pair of goggles.

4. *Method of Scoring:* The groups were plotted immediately by spotters in the pits on

reduced score sheets (1 inch equal to 1/16 inch). Each hit was plotted exactly in accordance with the location of the hit on the target by the score keeper. These score sheets provided the basic data for the analysis of the results. Not more than 5 per cent error inheres in this method of scoring.

After each day's firing, the data were analyzed according to the standard method known

Table 2
Firing Schedule

Relay	String	Subject						
		A1	A2	A3	A4	A5	A6	A7
1	1	None*	None	None	None	None	None	None
	2	A**	B**	C	D	E	F	G
2	3	None	None	None	None	None	None	None
	4	B	C	D	E	F	G	A
3	5	None	None	None	None	None	None	None
	6	C	D	E	F	G	A	B
4	7	None	None	None	None	None	None	None
	8	D	E	F	G	A	B	C
5	9	None	None	None	None	None	None	None
	10	E	F	G	A	B	C	D
6	11	None	None	None	None	None	None	None
	12	F	G	A	B	C	D	E
7	13	None	None	None	None	None	None	None
	14	G	A	B	C	D	E	F

* Without goggles.

** Goggle "A," goggle "B," etc.

as the Mean Radius Method. In this method the Group Center is located by a simple graphic process. The distance (linear in all cases) between the determined Group Center and the Center of the target is known as the Group Off Center measure. When the Group Center has been determined, the deviation in inches of each shot from this point is measured. The average deviation is known as the Mean Radius, and indicates the variability about this center. We then have a measure of central tendency (Group Center) and a measure of variability about this center. In the experiment, the best measure of firing accuracy is given by the Mean Radius score. The Group Off Center measure is subject to influence by constant errors such as improper sighting, wind shifts, etc.

The Method of Mean Radius is a method used to obtain a precise measure of the variability of a group of shots. It is based upon some simple statistical concepts. The method as used by riflemen is as follows:

Given the distribution of shots, two coordinates are drawn as follows: (1) a horizontal coordinate through the lowest shot in the group, and (2) a vertical coordinate through the shot nearest the left hand side of the target. From the horizontal coordinate, the distances are measured (in inches) between the coordinate (lowest shot) and each other shot. These distances are averaged, and a new horizontal coordinate drawn on the target as a dotted line. The same operation is performed with the vertical coordinate. When the new vertical (dotted) line is drawn in, the point of the intersection of the two dotted coordinates is known as the "Group Center."

From this Group Center, the shortest linear distance (in inches) between it and each of the shots is measured and recorded. The average distance is determined for each distribution of shots fired and is called the "Mean Radius."

It can be readily seen that this is a quick method of determining a measure of central tendency (the Group Center) and a measure of dispersion around that center (The Mean Radius).

Results

Basic Data: The basic data upon which all the calculations were made are not shown in this report. The Mean Radius score (in inches) for each string of shots under each testing condition for every subject was calculated for each day of firing. An inspection of Table 3 reveals that the average Mean

Table 3

Summary Table Showing Average Mean Radius Scores

Without Goggles	4.2
Goggle A	4.5
Goggle B	4.3
Goggle C	4.2
Goggle D	4.3
Goggle E	4.1
Goggle F	4.3
Goggle G	4.3

Radius, regardless of test condition, varied from 3.9 to 4.9 inches, and the standard deviation varied from 0.8 to 1.4. However, a study of the variability of the average Mean Radius does not indicate any significant trend.

Summary of Data on Firing Accuracy: The averages of the Mean Radius scores for each test condition (without goggles and with each type of goggle) summarized for all days of firing are presented in Table 3. The evidence suggests the conclusion that the wearing of any goggle tends to reduce firing accuracy. Except for goggles "E" and "C", the "without goggle" control firing was superior to firing with goggles. Hence, the control goggle "A" (clear plastic) is probably the best standard with which to compare the effect of other filters, rather than the "without goggle" condition. Whether or not the small differences obtained can be regarded as significant will be discussed in the following section.

Significance of Obtained Differences Among Test Goggles: In order to estimate the significance of the differences between shooting with various goggles and shooting without goggles, the following statistical test was made. The largest average Mean Radius score difference was taken from Table 3. This difference represented a magnitude of 0.4 inches in the average Mean Radius between goggle "E" (4.1) and goggle "A" (4.5). The Critical Ratio of this difference was found to be 1.85, or 96.5 chances in 100 that the obtained difference was a real one and not due to chance. This level of significance is below the one usually accepted in work of this kind. Thus, we have shown that even the largest of Mean Radius differences, though suggestive, is not statistically significant. It was therefore not considered necessary to investigate the other differences along these lines.

Daily Trends: In Figure 1 is shown the distribution of Mean Radius scores (without goggles) for the first, second, and third days of firing. In Figure 2 is shown the distribution of Mean Radius scores with goggles for the same three days of fire. It will be noticed that there is no significant day-to-day trend in either of the two figures, and that the distributions may be superimposed. It may therefore be concluded that the continued wearing of the various shooting goggles did not in any way significantly increase the marksmanship score.

Analysis of Variance: An analysis of variance was made, the results of which are summarized in Table 4. The following interpretation is offered. The differences among the goggles are clearly non-significant, and there is no evidence of day-to-day *general* shifts in firing accuracy.

The differences among the subjects are highly significant which is, of course, not surprising. There is no statistical significance in the goggles $\times$ subjects interaction. In other words there is no evidence that, for example, goggle A is

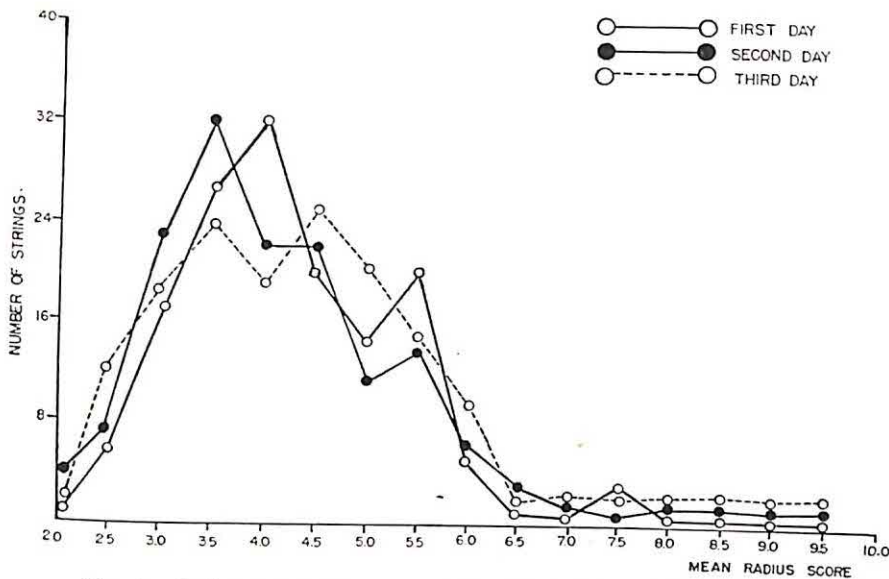


FIG. 1. Daily mean radius score of 21 subjects firing without goggles.

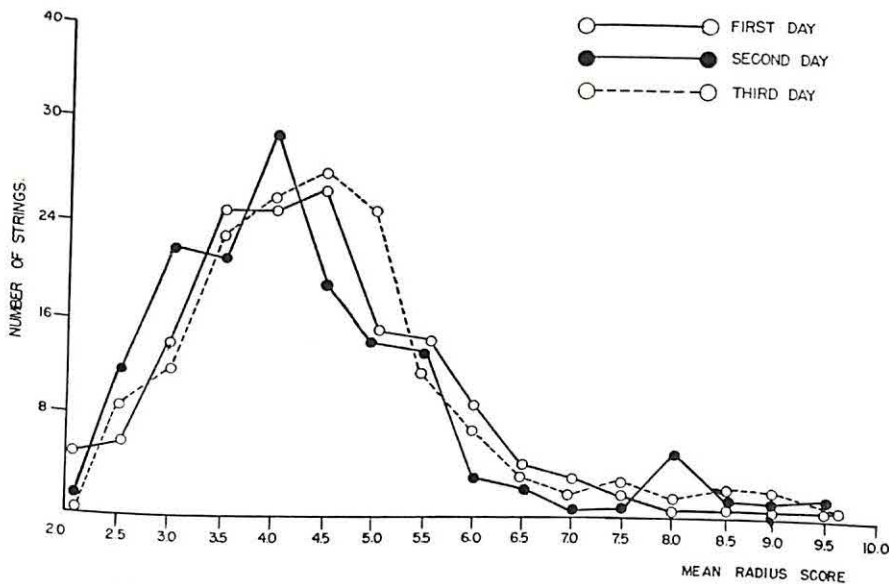


FIG. 2. Daily mean radius score of 21 subjects firing with goggles.

Table 4
Analysis of Variance

Source of Variance	Sums of Squares	Degrees of Freedom	Variance	F
1. Goggles	5.3602	6	.893	(non. sig.) F < 1.0
2. Days	1.2456	2	.623	(non. sig.) F < 1.0
3. Subjects	154.8697	20	7.743	8.27 (1% level)
4. Goggles $\times$ Subjects	128.2571	120	1.069	1.14 (non. sig.)
5. Days $\times$ Subjects	65.2087	40	1.630	1.74 (1% level)
6. Goggles $\times$ Days	33.3077	12	2.776	2.97 (1% level)
7. Error	224.7513	240	.936	
8. Total	613.0183	440		

better than goggle B for subject 1, but worse for subject 2, etc. The days $\times$ subjects interaction is highly significant and means merely that the relative standings of the subjects in firing performance is a function of the day on which performance is measured. In other words, subject 1 may be better than subject 2 on one day, but poorer on another day.²

The value for the goggles $\times$ days interaction is highly significant and poses a real problem. It must be recalled that the two simple variances, goggles and days were not significant. This may mean that one goggle is significantly better than another on one day, but not on another day. Over the period of days, the differences could cancel out.

Several factors might have varied during the test situation. For example, the conditions of light, haze, etc., might have varied slightly from day-to-day in ways which would affect the relative efficiency of the several goggles.

There may have been differences in the subjects which could have come into play. The preferences of the subjects might have in-

fluenced the day's firing. The subjects were interviewed in regard to their preferences, but the stated preferences bore no particular relationship to performance.

Summary

Six different types of plastic filters were tested as shooting glasses at the Rifle Range, Camp Lejeune, North Carolina. A neutral, clear plastic was used as a control. The subjects were 21 riflemen with the highest qualifications. They fired the standard M-1 rifle from the prone position at an "A" target from a distance of 200 yards.

The shots were plotted immediately on reduced score sheets of target design. The accuracy of fire was determined by the standard method, known as the Mean Radius method.

No statistically significant difference in performance, as determined by the Mean Radius scores, was found between firing with glasses and firing without glasses. No significant day-to-day trends were found.

The major conclusion reached was that the use of the plastic filters did not enhance range firing accuracy.

Received August 9, 1949.

² This is also to be expected, although it is a principle which athletic coaches are more familiar with (and possibly deal with more effectively) than experimental psychologists. There may be changes in general fitness from day to day, level of aspiration, etc.

Evaluation of Aircraft Instrument Displays for Use with the Omni-Directional Radio Range (VOR) *

A. C. Williams, Jr., and S. N. Roscoe

University of Illinois

The omni-directional radio range (VOR) provides the pilot with continuous visual information concerning his bearing to or from an omni-range station. This makes it possible to fly to a station along any desired track or from a station in any desired direction, representing a considerable advantage over the conventional four-leg auditory radio range. In addition, VOR employs a very high frequency signal which provides static-free reception. It is, however, limited to line of sight distance.

The purpose of the present study was to evaluate selected aircraft instrument displays which could be used with VOR. The evaluation was not based on engineering criteria, since the human engineer is not primarily concerned with the operation of the machine as such, but rather with the performance of the man operating the machine. Thus the dependent variables which were measured and compared were the speed and accuracy with which representative pilots could use the displays in question. An inspection of existing VOR instrument displays was sufficient to suggest that pilots would not be able to use these instruments either quickly or without frequent errors because of the inherent ambiguity in the immediate information presented.¹

The problem seemed essentially to be one of mode of display. Information presented by all displays was correct, adequate and clearly legible, but it was often difficult to interpret. For this study, three new displays were devised which, it was believed, might represent an im-

provement in ease of interpretation compared with existing displays. For comparison these three were evaluated in conjunction with five conventional displays of varying design. In all, eight displays were tested and compared. They are illustrated in Figures 1 to 8, inclusive.² They were: (1) Conventional Symbolic Indicator, (2) Air Line Indicator, (3) Air Force Indicator, (4) Experimental Symbolic Indicator, (5) Radio Magnetic Indicator, and (6), (7) and (8) Pictorial A, B and C, respectively, special displays designed for this study. For illustrative purposes Figures 1 and 6 only are presented here:

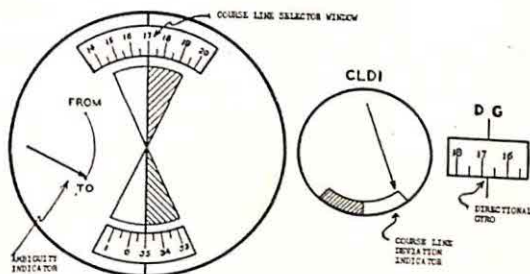


FIG. 1. The CONVENTIONAL SYMBOLIC INDICATOR. The Course Line Selector reads: your selected course is 170 degrees. The Ambiguity Indicator reads: your selected course is TO the station. (If you were South of the station, it would read "FROM," indicating that your selected course of 170 degrees radiated out from the station in that direction.) The Course Line Deviation Indicator reads: you are to the LEFT of your selected course. (Without the Directional Gyro reading of 168 degrees you would know only your approximate position and would have no way of knowing in which direction you were flying at the moment.)

* This article is based upon research conducted at the University of Illinois, Urbana, Illinois, under the auspices of the National Research Council Committee on Aviation Psychology, with funds provided by the Civil Aeronautics Administration.

¹ Melton, A. W. *Psychological problems in cockpit instrumentation for the omni-directional range (ODR) and distance measuring equipment (DME)*. Washington, D. C.: CAA Division of Research, Report No. 76, February, 1948.

² To reduce printing costs, Figures 1, 2, 3, 4, 5, 6, 7, and 8 and Tables 3, 4, 5, 6, 7, and 8 have been deposited with the American Documentation Institute and may be ordered as Document 2742 from American Documentation Institute, 1719 N Street, N.W., Washington 6, D. C., remitting 50¢ for microfilm (images 1 inch high on standard 35 mm. motion picture film) or 60¢ for photocopies (6 X 8 inches) readable without optical aid.

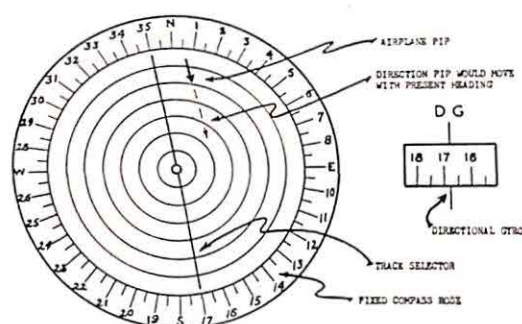


FIG. 6. PICTORIAL A. This instrument is a visual display representing an area about the VOR station. North is always at the top. The station is the small circle in the middle of the scope. Your airplane appears as an arrow-shaped pip which shows the actual spatial relationship of your airplane to the station. The arrow head of the pip shows the heading of your airplane which can also be checked by referring to the Directional Gyro. The Track Selector serves only as a reference line indicating your desired track to or from the station.

The information given by the eight displays shown is in each case the same. All indicate that the airplane's present position is somewhere to the North of the VOR station and that it is flying on a heading of 168 degrees. On the first four displays the airplane's position is inferred from the fact that it is somewhere to the left of a hypothetical course of 170 degrees to the station. On the others its azimuth position is indicated directly. On the pictorial displays the airplane's distance from the station is also shown, although this information is not given by the VOR signal and was not necessary for the solution of the problems used in the experiment. In all displays the airplane's heading is indicated directly.

For those unfamiliar with such displays, some idea of how they are used and of their relative difficulty can be gained by a simple exercise. In each case assume that the pilot's task is: *To fly to the station along a track of 170 degrees*, and then try to decide how he should fly in order to solve the problem. What is the first thing he should do? (In each case the pilot should first turn *right* to a heading of approximately 215 degrees. This would cause him to intercept his desired track at about a 45 degree angle, thus putting him in a position to turn to the heading of his desired track and fly directly to the station.)

Designs (1) to (5) represent primarily sym-

bolic modes of display, because they present information by means of numerical pointer readings, needle deflections, or numbers appearing in windows. Designs (6), (7) and (8) were intended to be graphic or pictorial displays because the actual horizontal spatial relations between aircraft, station and heading are shown in miniature on the face of the instrument. In general, symbolic displays provide pieces of information appropriate for a cognitive solution to a navigation problem, while pictorial displays provide cues appropriate for a perceptual solution to a problem. In this case the dichotomy was not absolute, because some of the symbolic displays also provided cues which could be interpreted perceptually, while the pictorial displays provided some symbolic quantitative information.

Considering these facts, an experiment was designed to measure and contrast the speed and accuracy with which representative pilots could use the eight displays in solving representative air navigation problems.

Description of the Experiment

The experiment was designed in such a way that variability in performance due to individual differences among pilots as well as differences among displays could be estimated by the analysis of variance technique. The design did not provide specifically for an accurate estimate of the possible interaction between the various displays and the different types of navigation problems for which they were used.

The displays evaluated were picture mockups instead of actual flight instruments. The mockups were made by drawing schematic pictures of the displays, the indications of which were made to read as they would on a real instrument had the aircraft been in a given position with respect to a VOR station. A series of ten navigation problems was made up based upon ten different positions of the aircraft with respect to the VOR station. The problems were selected so as to sample a variety of flight situations in which VOR might be used. Each problem specified a task which the pilot was supposed to solve by reading the display. The tasks in general were either to fly to or away from the VOR station along a track through the aircraft's present position.

or to fly to or away from the station on some other designated track as might be required by traffic control. To indicate his solution the pilot was given a double multiple choice of answers. On the first part he had to indicate whether he would turn right, fly straight, or turn left to initiate his solution. On the second part he had to choose one of four headings to which he would turn if in the first part he had decided to turn, or he could indicate he would maintain the same heading if he chose not to turn. For example, if his problem were: "To approach the station along a track through my present position," his double multiple choice of answers might be:

turn right "I should fly straight to/on a heading of turn left	North. 90 degrees. 180 degrees. 270 degrees."
--	--

All of the answer choices were in this form. The subject was required to circle the proper words in the first and second parts so as to make a complete statement of his solution of the problem.

A set of ten mockups representing the ten problems was drawn for each display. The problem sets were similar for all displays but not identical. The basic spatial relations in a given problem remained the same for all displays, but bearings, headings, and tracks were systematically varied from one display to the next so as to prevent pilots from memorizing the problems as they went from one problem set to the next. The order of problems within each set of ten was different for each display so as further to discourage learning the problems.

Subjects

The problem sets were given to 48 pilots divided into three groups as follows: Group I—16 non-instrument pilots with approximately 100 hours of flight experience; Group II—16 commercial pilots with CAA instrument ratings, most of whom were also flight instructors; Group III—16 scheduled airline pilots.³

³ The authors wish to express their appreciation to the 16 United Airline pilots who served as subjects and to Col. A. D. Tuttle and Dr. G. J. Kidera who recruited their services and arranged for the testing sessions.

Procedure

In order to introduce the pilot-subjects to their task, it was found necessary to prepare elaborate written instructions supplemented by verbal briefing concerning the displays and the types of problems to be solved. A set of general instructions covering all displays and problem types was issued. These were followed by a specific set of instructions for each display. These written instructions were too extensive to be included here. However, they described as clearly and completely as possible the interpretation and use of each part of each display and the specific flight procedures which would be followed in solving each type of problem.⁴

The problems and procedure were pretested using a separate group of eight subjects. As a result of the pretest it was found necessary to change some of the mockups because of previously undetected errors in their readings. Likewise it was found necessary to rewrite many parts of the instructions because they were not understood by the pre-test subjects. In general the instructions had to be amplified.

Experimental Design

Since all displays could not be used simultaneously by all pilots, it was necessary to arrange the sequences in which they were used by different pilots so as to balance the effects of practice, transfer and fatigue. Since there were eight displays, it was necessary for them to be used in eight different serial sequences. By virtue of such an arrangement, each display could be presented once and only once in each serial position while also appearing only once in each of the eight serial sequences. Two subjects from each group, or six subjects in all, were assigned to use the displays in an order according to each sequence.

How well displays could be used by the pilot-subjects was measured by the time required by a subject to complete each set of ten problems and also by the number of incorrect answers in each set. A problem was considered in-

⁴ Complete reproduction of these instructions may be found in Williams, A. C., Jr., and Roscoe, S. N. *Evaluation of aircraft instrument displays for use with the omni-directional radio range*. Washington, D. C.: CAA Division of Research, Report No. 84, March, 1949.

correctly solved if either part of the double answer were wrong.

Results

1. *Comparison of the Groups.* Before any data pertaining to differences among displays were analyzed, both the time and error scores for the three groups were tested for homogeneity by use of the *t*-test. The three groups did not differ significantly in their time scores. With respect to error scores the groups were different, both of the more experienced groups making significantly fewer errors than the non-instrument group. Because of this difference the three groups will be treated separately in the evaluation of the displays.

The speed and accuracy with which the various pilot groups used each display are shown in Table 1. The time scores represent the average time required to complete each set of ten problems. The error scores represent the average number of problems incorrectly solved per set of ten.

There was a strong relationship between the speed and accuracy with which the three groups used the various displays. This is shown by the rank order correlation coefficients listed in matrix form in Table 2. These correlations indicate that displays which were used more rapidly also tended to be used more accurately. This relationship obtained for all groups of subjects, even when the time scores for one group were correlated against the error scores for a different group.

2. *Comparison of the Displays.* Table 1 shows considerable variability among the time

and error scores for the different displays. For example, using Pictorial A, the airline pilots required on the average six minutes to solve ten problems, of which only 1.5 were wrong; whereas they required 21.1 minutes and made 4.8 errors per ten problems using the Air Force Indicator. Similar ranges were found for all three groups.

As a preliminary test to determine whether these apparent differences between displays were significant enough to warrant more detailed treatment, analyses of variance were made for the raw speed and error scores of the 16 subjects in each of the three groups. The design for these analyses was categorical by displays and subjects. Thus, variance *between displays* and variance *between subjects* were compared with the *residual* variance. Since error scores on individual problems were dichotomous, i.e., each problem was scored either right or wrong, and since time measures for individual problems were not obtained, it was not feasible to categorize the data in terms of problems. Summaries of the six analyses are given in Tables 3 to 8, inclusive, which are deposited in ADI (see footnote 2).

These six analyses show the results to be consistent for all three groups. There were in all cases significant amounts of variance (at the 1% level of confidence) attributable to differences among the displays. Practically all of the variance was accounted for in terms of two sources: (1) differences between displays and (2) differences between subjects.

On the basis of the indicated differences among the displays, individual comparisons

Table 1

Average Time and Error Scores per Set of Ten Problems for Each of the Eight Displays Tested
(Time scores are in minutes)

Display	Group I		Group II		Group III	
	Time	Errors	Time	Errors	Time	Errors
Pictorial A	7.9	2.8	5.6	0.9	6.0	1.5
Pictorial B	9.5	4.4	6.4	2.8	9.9	2.1
Pictorial C	10.0	4.8	8.5	2.8	10.4	2.5
Radio Magnetic Indicator	14.2	5.8	15.4	4.0	14.9	3.1
Air Line Indicator	17.2	6.3	15.4	4.1	15.9	4.1
Air Force Indicator	20.3	6.6	19.9	3.8	21.1	4.8
Experimental Symbolic	14.2	6.0	13.4	4.3	13.3	5.0
Conventional Symbolic	18.4	6.9	19.4	4.3	18.8	4.6

Table 2

Matrix of Rank Order Coefficients Between Time and Error Scores for Different Displays by the Three Subject Groups

Variables: T_I = time scores for Group I, E_I = error scores for Group I, T_{II} = time scores for Group II, etc.

	T_I	E_I	T_{II}	E_{II}	T_{III}	E_{III}
T_I						
E_I	.97					
T_{II}	.98	.94				
E_{II}	.72	.83	.68			
T_{III}	.99	.95	.99	.69		
E_{III}	.82	.83	.76	.80	.76	

were made by the t -test for correlated means. The error scores for each group on each display were pitted against the corresponding scores for each of the other displays. A similar treatment of the time scores was not made. However, similar differences would be expected to obtain in both cases, since the time and error scores were known to be highly correlated (as shown in Table 2). In any event, accuracy would seem to be a more critical criterion than speed in this particular case.

The results of these comparisons of the error frequencies for the different displays are shown in Table 9. The displays are listed at the left according to their rank order based on error scores. Across the top they are listed in the reverse order. Any given percentage appearing in the matrix refers to the level of significance of the superiority of the display indicated to the left over the display indicated above. The matrix could be constructed in this way because there were no reversals in the hierarchy of differences. Although a higher ranking display was not always *significantly* better than all lower ranking displays (see lower part of matrix), a lower ranking display was never superior, for any group of subjects, to one ranking above it.

Table 9 reveals the following results. There were virtually no significant differences among the five displays designated as symbolic. In four scattered cases the Radio Magnetic Indicator was used more accurately than the three lowest ranking displays. In three of these four cases this advantage obtained only for the airline pilots who may have had some additional

experience with some similar type of display, for example, the radio compass.

In general, this matrix emphasizes the dichotomy, discussed earlier, between symbolic and pictorial displays. Furthermore, pictorial displays should be centered about the station (with North at the top) rather than about the aircraft, as demonstrated by the superiority of Pictorial A over either Pictorial B or C. The airplane, rather than the station, should be represented by the moving part.

3. *Reliability and Internal Consistency of the Test.* Inasmuch as the experimental technique employed consisted of a paper and pencil test, such a test should meet certain standards of reliability and internal consistency.

No direct reliability data were available since a single form of the test was given only once to any one subject, and since the individual problem sets were hardly long enough to employ the split half method. However, the reliability of the test may be inferred to be high on the basis of the consistency of the results for the three independent groups of subjects (see especially Tables 1, 2, and 9). Furthermore, an item analysis showed the test to be internally consistent, since each of the ten test items or problems correlated acceptably with the total test error scores for each of the three groups. On the average the individual items correlated about .60 with the total test. (Time scores were not obtained for the individual items.)

4. *Dependence between Problems and Displays.* It would be of interest to determine whether some instruments are relatively more effective than others in working certain types of problems while being less effective for others, i.e., whether there was any "interaction" between instruments and problems. Unfortunately, however, the data are not such as to yield a direct answer to this question. Such interaction could not readily be evaluated by analysis of variance procedures for reasons stated previously, namely, error scores for individual problems were dichotomous and time scores for individual problems were not obtained. Nor were the assumptions required for the use of certain other statistics, such as chi-square, completely justified. Inspection of the raw error data suggested, however, that

Table 9

Matrix of Significant Differences in Error Scores for Groups I, II and III on the Different Displays

Percentages indicate the level of significance of the superiority of the display indicated at the left over the display indicated at the top, as determined by the *t*-test for correlated means. The single division line emphasizes the dichotomy between the "pictorial" and the "symbolic" displays. The double division line emphasizes the superiority of the station centered pictorial display over all others, including the airplane centered pictorials.

Composite Error Rank	Display	Group	C.S.I.	E.S.I.	A.F.I.	A.L.I.	R.M.I.	Pic C	Pic B	Pic A
1	Pic A	I	.1%	.1%	.1%	.1%	.1%	2%	2%	
		II	.1%	.1%	.1%	.1%	.1%	1%	1%	
		III	.1%	.1%	.1%	.1%	.1%	5%	—	
2	Pic B	I	.1%	5%	.1%	2%	2%	—		
		II	2%	1%	—	5%	1%	—		
		III	.1%	.1%	.1%	.1%	2%	—		
3	Pic C	I	1%	—	1%	2%	—			
		II	5%	5%	—	5%	—			
		III	1%	1%	1%	1%	—			
4	R.M.I.	I	5%	—	—	—				
		II	—	—	—	—				
		III	5%	1%	2%	—				
5	A.L.I.	I	—	—	—					
		II	—	—	—					
		III	—	—	—					
6	A.F.I.	I	—	—						
		II	—	—						
		III	—	—						
7	E.S.I.	I	—							
		II	—							
		III	—							
8	C.S.I.									

there were no marked trends indicating dependence between problems and displays. Since a wide variety of possible flight situations was sampled by the ten problems, this apparent lack of interaction is encouraging in that it further justifies making generalizations concerning the relative overall effectiveness of the various displays for air navigation.

Discussion of Results

Within the scope of this experiment the results have shown that pictorial type VOR displays can be used to solve navigation problems with greater speed and accuracy than can symbolic type VOR displays. Whether these results are applicable to flight depends upon the significance of the discrepancies between the

experimental task and the flight task. In the use of any such instrument display, the pilot's task involves two aspects of performance, one of discrimination in which the pilot must decide what to do, and one of manipulation in which he must execute his decision.

In this experiment, it was possible to get at only the first aspect of the pilot's task, namely: his decision concerning which way to fly in order to solve the navigation problem at hand. For this purpose it is believed that the results are pertinent to the flight situation. On the other hand, by using mockups some information was lost which could have been obtained in flight by watching the trend of the instrument readings. But this handicap would

apply equally to all displays and would probably not affect their relative standing.

A second disadvantage of mockups is that the subject could not manipulate the instrument so as, for example, to change a "to" course to a "from" course or vice versa, as could be done with some of the actual instruments. At the same time the opportunity to manipulate an instrument is also an opportunity to make an error in manipulation as has been found to be the case in a Link trainer equipped with VOR. Even though the inability to manipulate the mockup might handicap some of the symbolic displays, the same handicap existed for the pictorial displays in the case of their track selector which could not be manipulated on the mockup but could on the real instrument. For this reason it is doubtful that lack of manipulation discriminated unfairly between the various types of displays. On the whole it is the opinion of the investigators that so far as the pilot's initial task of orientation and discrimination is concerned the results of this experiment would recur in the actual flight situation.

After deciding what to do as a result of reading and interpreting a display, the pilot must then execute his decision by flying the airplane accordingly. To do this the pilot uses additional instruments such as the artificial horizon, altimeter, etc. But he also uses the VOR display as a source of direction information along with the compass and directional gyro. When used in this way the display can be considered to be a flight instrument and, as is true in the case of other flight instruments, the pilot's task then becomes chiefly one of alignment. Using the aircraft controls the pilot must manipulate the airplane so as to align a moving indicator of some sort with a previously determined fixed reference on the display.

This sort of performance cannot be evaluated by use of static mockups because there are no controls and no moving parts involved in the mockups. It would be unfair and unwise to suppose that the results of this experiment could be used to infer the adequacy of these displays in their capacity as flight instruments. To the contrary there is reason to believe on theoretical grounds that a more accurate flight path can be achieved using some of the sym-

bolic displays than by using the pictorials. Symbolic displays using the course line deviation indicator are more sensitive to displacement from the selected track than are the pictorial displays. Both the Air Force Indicator and the Experimental Symbolic Indicator were designed so that an asymptotic approach to a selected track could be achieved by maintaining a simple alignment on the face of the display. Presumably the ease of accomplishing this performance using these displays would not be surpassed using pictorial displays. Certainly an evaluation based on mockups is not relevant in this respect.

On the whole it is accurate to say that this experiment has thrown no light on the adequacy of the displays when they are used as flight instruments. There is reason to suspect that those displays which were found in the experiment to be superior would not necessarily be superior when used in this manner.

It must be remembered, however, that VOR is primarily a navigation device. Its chief purpose is to supply the pilot with information from which he can orient himself and from which he can decide on a proper direction in which to fly. If the VOR display fails in this respect so that the pilot makes wrong decisions, then whatever excellence it might have as a flight instrument is wasted. If it were a case of the pilot doing the wrong thing very well as opposed to, perhaps, doing the right thing passably well, the latter alternative is to be preferred. This experiment has contributed evidence bearing on the inclination of pilots to make correct or incorrect decisions as a result of using different kinds of VOR displays. For this reason it is felt that the experiment is pertinent to the problem as it exists in flight, and that the experimental results should be considered when making a choice of displays for actual use.

Summary

The speed and accuracy with which 48 pilots could use mockups of eight different VOR aircraft instrument displays to solve typical navigation problems were measured. The pilot group was composed of 16 non-instrument pilots, 16 commercial pilots with instrument rating, and 16 scheduled airline pilots. The instrument and airline pilots made fewer errors

than the non-instrument pilots but there was no significant difference in time scores. The rank order of displays based on error scores was highly correlated with their rank order based on time scores both within each pilot group and between groups.

With respect to both time and error scores there were significant differences between displays, similar differences being found for all groups. So-called pictorial displays which presented information in terms of a graphic representation of the actual spatial relations involved were significantly superior to the so-

called symbolic displays which presented information in terms of dial readings, needle deflections and numbers. One pictorial display was superior to all other displays.

The reliability of the techniques used was inferred to be adequate from measures of internal consistency and from the similarity of results obtained from independent groups. Inspection of the data suggested that no marked trends indicating dependence between problems and displays were evident.

Received July 8, 1949.

The Prediction of Persistency in Premium Payment

S. Rains Wallace, Jr., and Alfred G. Whitney

Life Insurance Agency Management Association, Hartford, Connecticut

This is a report of an attempt to develop methods for predicting life insurance premium paying behavior. While the research was initiated for a purely practical purpose, it offers findings of interest to the academic psychologist, since it demonstrates that better than chance prediction of a response which has a quite real social and economic meaning to the individual can be made over a period of five years on the basis of some simple and fragmentary background data.

The Problem

Most life insurance companies (the commonly held belief to the contrary) lose money if a policyholder fails to continue his premium payments for at least two or three years. Even subsequent terminations of the policy are costly to the individual agents and agency managers, and it can be demonstrated that a high policy termination rate in a company, other factors being equal, results in an unhealthy financial condition. Furthermore, the companies are not unaware that the problem has deeper sociological implications. The termination of a policy almost always means loss to the policyholder, both financial and psychological. It represents a failure in a plan for security and, to some degree, an implication that the company has been inadequate in meeting the problems and needs of its client. For these reasons, the companies are eager to find factors which are associated with persistency in premium payment and which could be employed in training their agents how to identify potentially persistent clients. Of course, efforts are also made by the companies to conserve policies already sold, but a previous study (3) has indicated that the persistency of a policy is much more susceptible to the factors which occur at the time of sale than it is to efforts made subsequently to keep it in force.

Since considerations of selling technique make the close catechization of a potential

buyer something less than desirable, it was necessary to limit the study to the information which is routinely obtained in the standard policy application form. Furthermore, some of this information is almost certainly unreliable (e.g., income, occupation, etc.) since it is based upon estimates or classifications made by the agents rather than actual records or the client's report. The problem, then, consists in determining methods for weighting these items of information in order to obtain the maximum degree of accuracy in predicting the eventual fate, i.e., the future persistency, of a policy at the time it is applied for.

Methods and Procedures

The Sample. The basic data for the study were first gathered in 1942 (2). Fifty-two life insurance companies supplied information on 12,499 Ordinary Life policies issued in May, 1942. Each company was assigned a quota of policies on the basis of the amount of new life insurance sold in 1940, selected in accordance with random sampling procedures. Various checks on certain aspects of the sample have supported the hypothesis that it may be regarded as typical of all Ordinary life insurance sold during May, 1942.

Five years later, the companies provided information on the subsequent history of almost all of these policies. One hundred twenty-seven of the cases were lost (every company contributed to the follow-up and the lost cases were distributed throughout the entire sample) so that the sample was reduced to 12,372. Of these, 1861 policies sold on the lives of persons under the age of 15 (so called "Juvenile" insurance) were removed. Among the adult policies, 1243 were refused by the client, thus producing a still-born transaction, while 154 cases were terminated by death, expiry, or maturation. These cases were also removed from the sample. The remaining 9114 cases included 3898 cases sold to men and 1245 cases

sold to women by Ordinary agents. There were 3971 cases sold by Combination agents.

For the purposes of this paper, only the study of the 5143 policies sold by Ordinary agents will be reported. However, it may be stated that similar methods were applied to the other group and similar results were obtained.

The Data. For the original sample, each company provided the following information for each case: (1) the insured's annual income (usually as estimated by the agent); (2) the amount of the policy; (3) whether the insured underwent a medical examination; (4) the type of policy (whole life, endowment, term, etc.); (5) the manner of premium payment chosen by the client (monthly, quarterly, semiannually, or annually); (6) the beneficiary (wife, children, parent, business partner, etc.); (7) the sex of the insured; (8) the age of the insured; (9) the marital status of the insured; (10) the occupation of the insured; (11) the amount of life insurance owned by the insured; and (12) the company or companies in which life insurance was owned.

For the follow-up sample, each company provided the following information for each case: (1) whether cash was received with the application; (2) whether the policy was issued as applied for or at some other rate, type, or amount; (3) the number of fractional years' premiums paid as of July, 1947; and (4) the status of the policy as of July, 1947 (in force, lapsed, surrendered, death, expiry, maturation, etc.).

The Analysis. Preliminary inspection of the data for males indicated that students should be treated as a special group. The elimination of them, in addition to those not gainfully employed and those who made single premium payments, reduced the sample to 3448.

For this remainder of 3448, a list of variables was studied and persistency rates

those: (1) who have higher incomes; (2) who choose to make annual premium payments; (3) who are older; (4) whose policies are issued on an examined basis; (5) who previously owned some life insurance; (6) who buy policies other than modified life or term policies; (7) who are proprietors, executives, or professional workers; (8) who buy policies of larger amounts; and (9) who are married.

Other variables which were studied and which exhibited a lesser degree of relationship are indicated below. There is a tendency toward a higher degree of persistency among those: (1) whose beneficiary is the wife or children; (2) who pay cash with the application; and (3) whose policy is issued as applied for.

The calculated persistency rates for this group are shown in Table 1. With respect to some variables the total is slightly less than 3448 because of missing data.

Since all of the factors found to be related to persistency are interrelated in varying degrees, the problem becomes one of "weighting"—in other words, of determining a method for combining nine variables in a manner to give the highest possible predictive effectiveness. If all of the divisions of the nine variables could be accurately expressed on numerical scales, this problem of weighting could be solved by standard statistical techniques. Since, however, it is difficult to quantify most of the variables, the relationship between each pair cannot be expressed simply.

All of the two-way relationships among the nine factors listed in Table 1 were therefore studied and tables indicating their nature were prepared.¹ Since these relationships could not be readily quantified, it was necessary to resort to the procedure of continuous subdivision of the groups and finally to study the persistency of individuals having certain patterns of the nine factors.

For the purpose of this subdividing, policies bought on a salary savings basis were eliminated, since it was believed that the experience from 1942 through 1947 with such policies would not be applicable to conditions in 1949.

The initial sorting of the data was made on

¹ The interested reader who wishes to examine these relationships in detail is invited to write to the authors for copies of these tables.

$$\left(\begin{array}{l} \text{Persistency Rate} \\ = \frac{\text{number of policies in force}}{\text{no. of pol. in force} + \text{lapses} + \text{vol. term.}} \end{array} \right)$$

were calculated for various subgroups. A summary of the chief relationships which were discovered is as follows. There is a tendency toward a higher degree of persistency among

the factors, income and mode of premium payment. An analysis made at that point showed that within income and mode of premium payment groups, the predictive power of amount of policy and marital status disappeared. Accordingly, these two variables were eliminated from further consideration. The seven remaining variables were each divided into a number of categories as shown in Table 2.

The total number of possible patterns under such a system of classification would be the product of the numbers of categories,— $4 \times 4 \times 2 \times 3 \times 2 \times 4 \times 6 = 4608$.

Table 1

Relation Between Certain Factors and Persistency:
Sales to Adult Males (excluding students)
by Ordinary Agents

Variable	Number of Cases	Persistency Rate (%)
Income		
\$5000 and over	415	81
\$3000-\$4999	590	75
\$1500-\$2999	1739	61
Under \$1500	704	49
	3448	64%
Mode of Premium Payment		
Annually	1250	75
Semiannually	594	65
Quarterly	1077	57
Monthly	318	55
Salary savings	209	37
	3448	64%
Medical Basis		
Examined	2400	70
Nonmedical	1038	50
	3438	64%
Previous Ownership of Insur.		
In same company	758	77
In some other company only	1445	66
None	1199	53
	3402	64%
Type of Policy		
Whole life	1009	66
Endowment and retirement income	699	65
Family income	318	65
Limited payment life	879	64
Modified life and term	529	56
	3434	64%

Table 1—Continued

Variable	Number of Cases	Persistency Rate (%)
Age		
45 and over	474	78
35 to 44	818	71
25 to 34	1376	65
Under 25	770	45
	3438	64%
Occupation		
Executives, proprietors, professional workers	665	82
Semi-executive and semi-professional workers	482	72
Agriculture and ranching	377	69
Clerical office workers	329	64
Sales clerks and salesmen	329	58
Factory and mine employees, skilled and unskilled labor, armed forces	1256	50
	3438	64%
Amount of policy		
\$6750 and over	363	75
\$3250-\$6749	734	69
\$2750-\$3249	173	63
\$2250-\$2749	368	61
\$1750-\$2249	488	60
\$1250-\$1749	235	57
Under \$1250	1077	60
	3438	64%
Marital Status		
Married	2656	67
Single	731	53
Widowed or divorced	51	59
	3438	64%

It was decided to make the final sort on occupation because it was believed that the data on this variable were probably the least reliable. Accordingly, by the time this sort was made, it was frequently found that much of the predictiveness of occupation had disappeared so that, at most, a classification into only two or three categories was required. For other factors, too, preliminary analysis often disclosed lessening of their influence at various points. For example, the differences in persistency rates among semiannual, quarterly, and monthly business disappeared in the highest income group and combination of

Table 2
Initial Variables and Categories

Variable	Number of Categories
Income	4
Mode of premium payment	4
Medical basis	2
Previous ownership of insurance	3
Type of policy	2
Age	4
Occupation	6

the data for these three modes of payment was possible. By such combinations, the total number of patterns was reduced from 4608 to a workable number with sufficient cases for study in each pattern.

In each of the separate patterns, the per cent of policies in force was calculated and the per cents were then converted to a single digit numerical rating using the equivalents shown in Table 3.

The data were now examined to determine fundamental trends and where inconsistencies in these trends could be ascribed to chance, the data were smoothed in accordance with the trend. A description of the method of performing this smoothing can be obtained by request to the authors.

Results

This analysis, then, resulted in a *Persistency Rater* with which a score could be assigned to each case.² The cases in the study were then scored and the actual persistency determined for each rating group with the results shown in Table 4. The biserial coefficient of correlation between the Persistency Rating and the actual five year persistency is shown for each group.

It must be remembered, of course, that such "validation" of the *Persistency Rater* is fallacious since it is based upon the same cases with which the weights were determined. Some shrinkage should be expected when the Rater is applied to another sample. However,

² A correction was made in the *Rater* for the changes in income which have occurred since 1942. This correction was based upon comparative frequency distributions of incomes among consumer units in 1942 and 1948 and upon the comparative values of the dollar in 1942 and 1949. The *Rater*, which is now being used by a number of companies, may be procured upon request to the authors.

the writers expect such shrinkage to be slight in light of the objective nature of the factors used. This expectancy has been considerably reinforced by the results of a recent study made by one life insurance company.

This company selected, at random, 489 cases sold in the first six months of 1946 and studied their persistency for a two-year period. The sample is not strictly comparable to the larger one here reported since it comes from a different year and since the over-all persistency of the group is considerably higher (82% instead of 66%). However, when these cases were scored on the *Persistency Rater*, the results, as shown in Table 5, indicated a very satisfactory level of validity.

Discussion

It is not believed necessary to labor the point that the discontinuance of life insurance payments is a response which has many implications for the individual in question. Equally apparent would seem to be the possibility that such a response might have considerable significance to the psychologist, particularly in the social field. While the motivation behind the buying and subsequent payment of an insurance policy is certainly something less than clear or simple, it does appear to involve some aspects of "altruism," of "social responsibility" and of "economic stability." This study demonstrates that the response itself must have considerable reliability even despite the fact that we may identify numerous factors which might affect it and which are largely inde-

Table 3
Conversion Table for Persistency Rates and Numerical Ratings

Per Cent in Force	Numerical Rating
100	9
90-99	8
80-89	7
70-79	6
58-69	5
48-57	4
38-47	3
28-37	2
18-27	1
Under 18	0

Table 4

Proportion of Placed Policies Which Remained in Force for Five Years and Persistency Rating

Persistency Rating	All Men Except Students		Male Students		Women	
	No.	Per Cent in Force	No.	Per Cent in Force	No.	Per Cent in Force
7 and over	808	85%	155	82%	483	86%
6	781	74	104	70	452	75
5	736	62	5	20	91	62
4	423	52	0	—	73	52
3 and under	461	35	0	—	16	38
Total	3209	66%	264	76%	1115	77%
	$r_{bis} = .46$		$r_{bis} = .30$		$r_{bis} = .36$	

pendent of the individual (economic readjustments, changes in marital status, job changes, etc.). The response would, therefore, appear to offer considerable promise of providing a criterion for investigations in the fields of personality or social psychology. While many of the factors used in predicting the response may be regarded as reflecting the transaction rather than the buyer's characteristics, they need not be so interpreted. For example, the fact that policies paid for on an annual basis are more persistent may indicate the superiority of this method of payment, but they may indicate, instead, that persons who are capable of or who choose such a method are more likely to make the desired response. The latter hypothesis gains some measure of support from the fact that persons who change from an annual premium payment to a more frequent one have almost as high a persistency as

those who continue the annual mode of payment (1). Here, apparently, the interest of the individual and willingness to go through the process of changing his mode of payment outweighs the fact that he has more frequent opportunities to make the undesired response.

A factor which could not be included in this study but which other studies (3) have shown to have considerable influence is the agent who makes the sale. It is known that policies sold by some agents have a consistently high persistency, while those sold by others are consistently low. While it seems probable that some of this effect is related to the agent's selection of clients, it appears that some characteristics of the relationship established between the agent and his client may also be important determiners of the client's future responses. It is in this field that the prospects for improving accuracy of prediction and for enhancing our understanding of the determining factors would appear to lie. Studies in this direction are now in process.

Received January 16, 1950.

Early publication.

Table 5

Proportion of Placed Policies Which Remained in Force for Two Years and Persistency Rating

Sales by One Company to All Men Except Students		
Persistency Rating	Number	Per Cent in Force
8	68	100%
7	142	91
6	103	81
5	78	75
4	64	71
3 and under	34	50
Total	489	82%
	$r_{bis} = .47$	

References

1. Life Insurance Agency Management Association. *Persistency 1942-1947*. Hartford, Conn.: Life Insurance Agency Management Assoc., 1949, pp. 123.
2. Life Insurance Sales Research Bureau. *The 1942 Buyer*. Hartford, Conn.: Life Insurance Sales Research Bureau, 1942, pp. 79.
3. Wallace, S. R., Jr., and Twichell, C. M. Factors affecting the persistency of "Orphan Business." *J. Amer. Soc. Chart. Life Under.*, 1948, 2, 398-408.

Book Reviews

Super, Donald E. *Appraising vocational fitness*. New York: Harper and Brothers, 1949. Pp. xiii+727. \$6.00.

The primary purpose of the author is to provide an objective and detailed evaluation of a number of the most important tests employed by vocational psychologists. This has been realized in a series of chapters which are thorough, straightforward, and with few exceptions free from ambiguity. The author has summarized for each test the most significant research, discussed it in terms of current theory and practice in mental testing, and evaluated it on the basis of its use and reputation in educational guidance, consultation clinics, and business and industry. The work is scholarly.

A second objective of the book is to provide the reader with work habits and thought processes which may enable him to evaluate new testing instruments and new research. The realization of the second objective does not appear to be as complete as the realization of the primary objective. Although Dr. Super has certainly not limited himself to the cataloging of tests and their essential statistics in the manner which is so prevalent among texts in this field, he does not make explicit many of the important concepts and psychological assumptions which are implicit in the evaluation of human behavior. He tends to avoid the discussion of controversial topics and does not strive to illuminate for the student the broad psychological basis for and general relevance of the samples of behavior elicited by tests. His discussion of the basic psychological significance of some of the test classifications is weak and his appendix on the fundamental quantitative principles of mental measurement is not sufficient to its purpose. Dr. Super's work is critical and intelligent but his discussion is no more profound than his sources.

The implications of a test score are usually discussed in terms of an average relation with some isolated criterion and (from the standpoint of most clinicians) with insufficient regard consistently manifested for such variable factors as peculiarities of the individual's motivational pattern, his characteristic reaction to

competition and the probable future enrichment or impoverishment of his environment. Dr. Super's insightful and stimulating discussion of these considerations is familiar to those who have read his well-known *Dynamics of Vocational Adjustment*. In his new book, however, Dr Super is primarily concerned with the portion of the total test variance which is related to some known criterion. The possible determiners of the residual variance (that which makes for errors in predictions) receive relatively little attention. The author's privilege to make this differential emphasis would require no justification and would draw no comment had he not in the closing chapters of his book discussed the techniques of vocational guidance and the conversational devices whereby tests results are to be presented to the individual.

The book's greatest value lies in its careful delineation of the most important respects in which a test may be used. The primary emphasis is on the use of tests. Only the most obvious or experimentally illustrated misuses of tests are emphasized; the subtle conditions under which tests may be misused are not strongly illuminated. In presenting his sources, the writer's usual critical discernment was on a very few occasions in abeyance; and although the work is not free from minor errors, these are few and do not weaken the book.

In general, Dr. Super's book comprises a thorough adherence to the literature and he has presented its implications for many important uses of tests in an expert manner. *Appraising Vocational Fitness* will be a very desirable required text for courses in measurements, individual differences, guidance, and vocational psychology.

J. R. Wittenborn

Yale University

Bellows, Roger M. *Psychology of personnel in business and industry*. New York: Prentice-Hall, Inc., 1949. Pp. xii+499. \$4.50.

This book by an industrial psychologist attempts to "clarify the boundaries of the personnel methods and management field from

the socio-psychological point of view." It has been written "for those who are interested in improving personnel management by use of personnel systems and procedures."

Part I—Development of Personnel Technology defines the goals of personnel management and describes historical influences leading to modern personnel methods. Part II—Tools for Effective Use of Personnel includes separate chapters on most of the standard topics found in personnel management texts, viz., criteria, job analysis, recruitment, interviewing, testing, training, job evaluation, incentives, merit rating, and turnover. Part III—Worker Satisfaction Through Human Understanding defines the field of "industrial social psychology" and discusses the need for and techniques used in employee counseling, communication among employee levels, attitude surveys, and suggestion systems. Part IV—Implications of Personnel Technology describes the rising professionalization and specialization of personnel management and the need for trained leaders and summarizes current trends in personnel practices as influenced by research in the field. Appendix A presents lists of publishers, organizations, publications, and schools as sources of additional information. Appendix B presents the Taylor-Russell tables. Each chapter ends with a list of selected references.

Both the commonly-referred-to studies and recent literature on the topics discussed are fairly well covered. Consideration of the human element is stressed throughout and due recognition is made of the importance of the social influences impinging upon the individual worker. Topics frequently neglected are pointed up, as evidenced by separate chapters devoted to recruitment, employee suggestion systems, and methods of improving communication between management and employees. The need for basic research and for considering the interaction among the various personnel techniques is emphasized throughout. That personnel psychology is on its way to becoming a profession is evident to the reader.

Unfortunately, the essentially good "content" of the book is not presented in an equally effective manner. The writing is redundant, tautological, uneven in difficulty, and sometimes even ungrammatical. Words and

phrases are frequently so loosely used as to become almost meaningless. Simple, painfully obvious comments are couched in involved sentences. Even straightforward statements are frequently unnecessarily qualified. The following are selected from literally scores of possible examples:

"Since Pile A represented the most favorable attitude toward the church, this represents a fairly favorable attitude toward the church." (p. 376)

". . . the job-analysis approach can be useful only if job analyses are made and. . . ." (p. 426)

"Managements, in considering whether non-directive counseling is profitable, are concerned with the problem of evaluating it." (p. 323)

"Systematic use of . . . are often overlooked." (p. 159)

"Accept his feelings and attitudes and reflect it." (p. 320)

". . . qualifications . . . is. . . ." (p. 326)

"The non-directive approach is probably best adapted to the severity of the maladjustment." (p. 323)

". . . success in studying and becoming an accountant." (p. 140)

"The sample had been picked . . . into one of five major criterion groups." (p. 153)

"It is a project of some size to arrange for all employees to meet in a single room, particularly in large companies." (p. 333)

"Luncheon menus . . . might best be placed in the restaurant or in the hallways leading to the restaurant. . . . Departmental information might be placed in the hallways that are used by employees in a given department." (pp. 337-338)

". . . differences . . . has." (p. 351)

"Indirect media for sharing information consist of results as a by-product of the several personnel methods that are discussed in this book." (p. 345)

"This way of doing does not work so competently in industry." (p. 437)

"[This test terminology] was found useful in designating the different types of predictor devices since in this program considerable value was found in the use of non-test predictors, which were systematically used in

the same batteries as test predictors, when found to increase the validities of such batteries." (p. 156)

"A list of 769 words prepared by a professor named Dale. . . ." (p. 355)

Possibly the author was trying to "write down" the material to a lower level of understanding than this reviewer assumes necessary. Perhaps he himself was trying to apply the Flesch techniques which he (rightly) recommends for use in employee communications. If so, this reviewer appreciates the effort but still considers the level of presentation as vague and wordy rather than clear and precise.

The usual minor errors and omissions found in first editions are present. Figure 1 is incompletely labelled. In Figure 25 the term "applicants" should be "employees." On page 132, "selection ratio" is improperly used, although defined accurately in Appendix B. Appendix B, incidentally, is referred to in the text merely in a footnote. Figure 42 is an essentially meaningless figure purporting to show "The Three Degrees" of social distance between worker and boss. Table 29 is presented with no reference to it in the text. Table 40 should be numbered Table 39. The format of Appendix A is such that it is quite difficult to separate one item from another in the long lists presented. Many of the "Selected References" at the end of the chapters do not indicate the relevant section of the reference. The addresses of several of the journals listed in Appendix A are incorrect or out-of-date.

It is easy for reviewers to point out omissions of content. As the field of personnel psychology expands, authors are being forced to select what materials to include. This reviewer was disappointed, however, in not finding reference to British and French industrial psychologists, to the use of experimental groups hired without regard to test scores as an important step in test validation research, to the work of the Civil Service Commission on job Classification, to the effect of the use of employment tests on quality of applicant, to the types of tests which have been found useful in predicting job success, to the T. W. I. program of training supervisors, to the use of non-financial incentives, to data on voluntary restriction of output by workers, and to the

Labor-Management Committees of World War II. In addition, although scattered references are made, there is no systematic discussion of accidents and their prevention, of monotony, of music in industry, of fatigue and efficiency.

This reviewer's over-all judgment is that the author has a potentially good book at hand but that it needs a careful going-over and condensing before it is worth much more than a casual reading of Chapter V. Attracting Personnel, Chapter XVI. Techniques for Improving Communications, and Chapter XVIII. Employee Suggestion Systems.

Albert S. Thompson

Teachers College,
Columbia University

Prentice-Hall, Inc. *The new cure for white collar unrest.* 1948. Pp. 48.

This forty-eight page pamphlet is written in the breezy style typical of Prentice-Hall reports and services for businessmen. It contains an introduction and four sections titled: what makes white collar workers tick, why white collar workers kick, how unions sell themselves to white collar workers, and what employers can do for white collar workers. Over half of the booklet is devoted to the last of these sections and covers such diverse matters as merit rating, suggestion systems, personal consideration, company rules, earmarks of a successful supervisor, job evaluation, pension plans, getting information to employees, morale surveys, etc. Inasmuch as this itemization includes only a small number of the many items covered, it is obvious that treatment must be at a superficial level. This reviewer could find nothing to warrant inclusion of "new cure" in the title. Although the booklet is not free from factual error and unsupported statements, it is basically accurate at a simple level. It might thus be of value to the "busy executive" who wishes to cover a maximum of the subject matter in a minimum of time. The pamphlet obviously is not intended for personnel psychologists or professionally trained personnel and industrial relations directors.

Clifford E. Jurgensen

Minneapolis Gas Company

Gilbert, Jeanne G., and Weitz, Robert D. *Psychology for the profession of nursing*. New York: The Ronald Press Company, 1949. Pp. x+275. \$5.00.

The authors state that this textbook has been prepared specifically for student nurses. The purpose is to present principles of psychology in such a way that the nurse can make use of them in understanding the patient and in making a successful personal adjustment to life and its problems.

The book is divided into three parts. The emphasis in Part I, Fundamental Principles of Psychology, is upon individual human behavior and the factors which determine its development. In Part II, Personality, Mental Hygiene and the Normal Patient, the integration of all traits in the development of personality is discussed; and special problems which have to do with the care of children, aged, and chronic and convalescent patients are presented. Part III, Personality Maladjustments and the Abnormal Patient, includes information about psychopaths and other wayward types, behavior disorders in feeble-minded and organically diseased patients, the psychoneuroses and the psychoses. Case studies are presented to illustrate various types of abnormalities. The final chapter includes a brief presentation of diagnostic procedures and therapeutic techniques.

In writing this particular text the authors have probably attempted to meet the needs of those schools of nursing which, up to the present time, have tried to cover the entire field of psychology in an introductory course usually given to students during their first year in the school. It is not adapted to the needs of schools which are attempting to integrate psychological principles into every clinical subject which is taught, and which, in addition, are providing experience in psychiatric nursing for all of their students. The material on fundamental principles of psychology, integration of personality, diagnostic tests and therapeutic techniques; which should probably receive greatest emphasis in a basic course in psychology; is so briefly presented as to be of questionable value unless supplemented with readings in other texts and periodicals. Yet, the references at the end of each chapter include only the names of entire books and

periodicals. Without more specific references to guide her, the student in nursing is unlikely to do extensive supplementary reading.

In the discussion of adjustment the student is warned that she will be expected to adjust to a regimented, restricted environment where good judgment and adult behavior are expected. She is told that nursing is hard and that she must be equipped physically, mentally, and emotionally to meet the hardship. A more positive approach, in which the student is challenged to think of nursing as an interesting and worthwhile profession, would, in the long run, probably have a better psychological effect than the negative emphasis upon the importance of adjusting to conditions as they are.

Helen Nahm

Duke University

Hovland, C. I., Lumsdaine, A. A., and Sheffield, F. D. *Experiments on mass communication. Studies in Social Psychology in World War II, Vol. III*. Princeton: Princeton Univ. Press, 1949. Pp. viii+345. \$5.00.

This is the third in a series of four volumes reporting the work of the Research Branch of the Army's Information and Education Division during the war. Volumes I and II, previously reviewed in this journal (33, 609-611), discussed the attitude surveys relating to the soldier's adjustment to army life, and to combat and its aftermath, carried out by the Survey Section of the Research Branch. The present volume reports the work of the Experimental Section which had as its primary responsibility experimental studies on mass communication.

An introductory chapter outlines the variables investigated and the general nature of the film research done by the Experimental Section. A basic distinction is "made between studies where the purpose is to evaluate a completed product and those where the purpose is to investigate variables by controlled variation" (p. 4). In terms of this distinction the book is organized into two sections, the first reporting the evaluative studies and the second the studies employing controlled variation. "In both kinds of studies the main emphasis was on the measurement of changes in knowl-

edge, opinion, or behavior produced by a film or other communication device" (p. 5).

The first three chapters of Part I are devoted to the evaluation of the "Why We Fight" series of indoctrination films. Chapter 5 discusses several evaluative studies involving an experimental comparison of alternative methods of presenting the same material, e. g., a sound motion picture and a film strip. Chapter 6 analyzes the effects of films on men from different intellectual levels.

In Part II, one chapter is devoted to a study of the short-time and long-time effects of an orientation film upon factual knowledge and opinions, another to a study of the effects on opinion change of presenting one side as compared with both sides of a controversial issue, and a third to the effect of audience participation during a film strip presentation on learning the phonetic alphabet. Chapter 10 is an extremely well done bit of writing which neatly summarizes and evaluates the work of the Experimental Section. Four important appendices follow which deal with some of the measurement problems encountered by the Experimental Section in their various studies.

The results reported, with one exception, will not be particularly revealing to psychologists. We find, for example, that it is possible to increase factual knowledge as a result of being exposed to films. Some effects are observed on opinions specifically covered by the films, but these effects are not as great as those on factual knowledge. Still less is the effect of films on opinions of a more general nature—opinions which the films were designed to influence, but which were not specifically covered by the content. We find also that those with greater intellectual ability learn more from a film than those with less ability; that announcing in advance that a quiz will be given on the content of a film facilitates learning; that audience participation also facilitates learning. These results, and most of the others not mentioned here, would be predicted in terms of current psychological knowledge.

The exception, mentioned above, concerns the prevalent notion among social psychologists that the influence of "propaganda" is merely the reinforcing of existing beliefs. But in a radio presentation study the Experimental Section found that "whether a man was

initially for or against the stand taken in the communication, his opinion tended to be influenced in the direction of *more acceptance* of the point of view argued for in the communication. In no case was there found a significant change opposite to the intended effect of the communication among those initially opposed, and in all cases where the audience as a whole showed a positive change, the change was positive among those initially opposed" (p. 269).

Although the concept of "attitude" was frequently used in the first two volumes, the authors of the present volume avoid this concept, preferring to state their discussion in terms of "opinions." "This usage reflects in part the fact that no satisfactory way was found of distinguishing between attitude questions and questions of opinion" (p. 265). But statements of opinion are distinguished from statements of facts in that "opinions are *interpretations* of available facts, which interpretations are difficult to verify or disprove directly" (p. 265). "Informed" opinions are those based upon "valid" interpretations of available facts. "Valid" interpretations, in turn, "are those which can be shown to be positively correlated with an index of ability to make valid interpretations" (pp. 275-276), for example, educational level. "Invalid" interpretations are those negatively correlated with this index. This distinction leads to the rather interesting definition of "propaganda" as "attempts to foster 'invalid' interpretations" (p. 276).

That there are difficulties with this analysis will be obvious if it is applied to the 1948 election campaign. Take the statement: "Mr. Truman has little or no chance of being elected." Would agreement or disagreement with this statement have represented "informed" opinion? But which would have been the "valid" interpretation of the available facts? Remember that a "valid" interpretation is one which is correlated with an index of ability to make "valid" interpretations and that one such index used by the Experimental Section was educational level. And now ask and answer the question as to the relationship between political party affiliation and educational level, and you should also have the answer to the question of whether it was the Democrats or the Republicans that engaged in "propaganda" during the campaign.

Many of the results of the studies are reported in terms of the average *per cent* correct answers rather than in terms of the simple arithmetic average. This can be somewhat misleading. For example, Figure 5, page 143, reports the average *per cent* correct answers for a control and various experimental groups. Taking only the figures for the control group and the movie group, we find that the former averaged 39.5% correct answers and the movie group 46.6%, giving a difference, apparently the result of the movie, of 7.1%. When we realize that the test consisted of only 39 items, this difference is not as imposing as the *per cent* figures make it appear. The test mean score for the control group would be 15.4 and for the movie group 18.2, giving a mean difference of 2.8.

It is impossible to do justice to this book in a brief review, but let me say that I found it highly provocative and stimulating. Few experimental studies dealing with such complex situations have been as carefully planned, executed, and analyzed as those reported in this volume. The results, and particularly the methodology, should prove of interest to psychologists, film producers, and anyone concerned with audio-visual aids in indoctrination or training programs—whether in education or industry.

Nor is the book concerned with just methodology and the reporting of experimental results. These authors have not been afraid to speculate, to hypothesize, to indulge in theory—this provides stimulating reading in an area where theory is largely lacking. Questions are not only answered but raised, with the result that sufficient research problems are suggested to occupy workers in this field for some time to come.

Allen L. Edwards

The University of Washington

Mathewson, Robert H. *Guidance policy and practice*. New York: Harper and Brothers, 1949. Pp. 294. \$3.00.

The growing pains and self-examination which have characterized the profession of psychology during the years since World War II have been characteristic, also, of the kindred and partly overlapping fields of guidance.

In the American Psychological Association, various divisions have been struggling with definitions and attempting to delimit fields, with both clinical and educational psychologists feeling that guidance belongs to them while the guidance psychologists contend that they represent a distinct field. So, also, in the broader fields of education and social service, the professions of psychology, education, and social work have been staking claims to guidance, each claiming it as its own. To complicate matters, there has for some years been considerable debate among specialists in guidance as to just how the term is to be defined, just what needs it attempts to meet, and just what services it should provide in order to meet them.

Guidance Policy and Practice is an attempt to define terms, clarify objectives, examine methods in the light of these objectives, and so delimit the field. It represents the thinking of a guidance specialist who has had long experience as a practitioner at both adolescent and adult levels (under educational auspices), and who has, during the past few years, given considerable time to the training of graduate students. It is at a strategic time in the development of guidance and of professional psychology that Mathewson has put his thinking on record.

In Part I Mathewson discusses fundamental factors such as psychological and philosophical concepts, individual and social needs, institutional settings, the psychology of the participants, and costs. Part II treats the implementation of policies in schools and in the community. Part III takes up current issues such as the scope of guidance, the responsibility of education, and the training and role of the counselor. Part IV examines the future, particularly the trend toward unity or toward a broader concept of guidance than has prevailed in some circles, educational institutions as the primary medium of guidance, and the development of sound national policy in a period of increasing federal activity.

Although the book is uneven (the chapter on *The Institutional Setting* seems to deal with the obvious, while that on *The Role of the Counselor* is a timely and effective exposition of a much more mature approach to counseling than, for example, either Williamson's or

Rogers'), it is on the whole thoughtful, penetrating, and constructive. This reviewer would like to focus on two major contributions, and to point out some related weaknesses.

Perhaps the strongest aspect of Mathewson's philosophy of guidance is his reiteration of the *developmental* nature of guidance. He sees it as a process of orientation of the individual to society and of the development of a mature, independent, socialized, self. This approach resembles that of many educators, and stands in clear opposition to and contrast with that of many guidance specialists who came to this field from social work, and who view guidance as a service for "persons in need . . . experiencing some breakdown in their capacity to cope unaided with their own affairs," whether these be the choice of a college major or rehabilitation for employment after injury. Mathewson conceives of *guidance as primarily educational and preventive* rather than remedial; he concludes that schools and colleges, as institutions with principal responsibility for education and development, are the major guidance agencies. He recognizes that guidance also has an adjustive or remedial function, but one of the defects of the argument (surprising in an author who has worked at the adult level) is the failure to recognize the fact that *guidance needs and services do not cease at age 22*. In the vocational sphere, for example, guidance does not actually cease with "early adjustment" (p. 120), but, as gerontology is increasingly demonstrating, continues through life, with new problems arising and new services called for at each of the life stages. In this reviewer's opinion, then, Mathewson's definition of the scope and agencies of guidance is too limited, even though his emphasis is sound.

The second contribution singled out for comment in this review is the discussion of the trend toward unity in the field of guidance. As the recommendations of policy committees of the National Vocational Guidance Association in the '30s and again last year have shown, and

as evidenced further by current committee work in the Council of Guidance and Personnel Associations, there is an increasing awareness of the fact that adjustment involves a total personality, that the problem of vocational choice, for example, is very much a part of the problem of developing a self-concept which is in harmony with reality. Mathewson recognizes the *unity of personality*, attempts to define guidance in broad terms, and forecasts increasing acceptance of this point of view. But in this reviewer's judgment he makes two mistakes made by many educationists working in the field of guidance: he fails to recognize sufficiently the *complexity of personality*, and, failing to do this, errs further by (perhaps unintentionally) leaving the reader with the impression that all the vocational or educational counselor needs to do in order to become a personal counselor is to recognize the unity of personality. To recognize the unity of personality and of guidance without recognizing the complexity of personality and of society, including the world of work, is to encourage amateurish work and to lower standards. It encourages vocational and educational counselors to venture into psychotherapy without adequate training in that specialty; it encourages clinical psychologists to venture into vocational and educational counseling without sufficient knowledge of opportunities and requirements in schools, colleges, and the fields of employment.

Despite these defects, Mathewson's little volume is a thoughtful and important contribution to thinking on the nature, functions, and organization of guidance. While it may not contain the final solutions to our current problems of definition, concern, and jurisdiction, it should be read by all who are interested in finding a truly constructive way out of our present functional and organizational dilemmas.

Donald E. Super

Teachers College,
Columbia University

New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor,
Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota

- Interaction process analysis.* Robert F. Bales. Cambridge: Addison-Wesley Press, Inc., 1950. Pp. 203. \$6.00.
- Giant brains or machines that think.* Edmund Callis Berkeley. New York: John Wiley and Sons, Inc., 1949. Pp. 255. \$4.00.
- Problems in personnel administration.* Richard P. Calhoun. New York: Harper and Brothers, 1949. Pp. 540. \$4.00.
- Immediate and retention effects of interpolated rest periods on learning performance.* Bertram Epstein. New York: Bureau of Publications, Teachers College, Columbia University, 1949. Pp. 77. \$2.10.
- Scholarships, fellowships and loans.* S. Norman Feingold. Boston: Bellman Publishing Co., Inc., 1949. Pp. 256. \$6.00.
- The expanding role of government and labor in the American economy.* Waldo E. Fisher. Bulletin No. 18. Pasadena: Industrial Relations Section, California Institute of Technology, 1950. Pp. 26. \$1.00.
- Survey of selected personnel practices in Los Angeles County as of April 1, 1949.* Bulletin No. 17. Robert D. Gray et al. Pasadena: Industrial Relations Section, California Institute of Technology, 1949. Pp. 74. \$2.50.
- Selected personnel practices of large employers in Los Angeles County as of April 1, 1949.* Circular No. 18. Robert D. Gray et al. Pasadena: Industrial Relations Section, California Institute of Technology, 1949. Pp. 12. \$1.00.
- Personality, development and assessment.* Charles M. Harsh and H. G. Schrickel. New York: The Ronald Press Co., 1950. Pp. 518. \$5.00.
- Psychology in everyday living.* Ralph Leslie Johns. New York: Harper and Brothers, 1950. Pp. 464. \$3.50.
- Practical job evaluation.* Philip W. Jones. New York: John Wiley and Sons, Inc., 1948. Pp. 304. \$4.00.
- Psychology of labor-management relations.* Arthur Kornhauser, Editor. Champaign, Ill.: Industrial Relations Research Association, 1949. Pp. 122. \$1.50.
- Practical and theoretical aspects of psychoanalysis.* Lawrence S. Kubie. New York: International Universities Press, 1950. Pp. 252. \$4.00.
- Rehearsal for destruction.* Paul Massing. New York: Harper and Brothers, 1949. Pp. 341. \$4.00.
- Patterns of panic.* Joost A. M. Meerloo. New York: International Universities Press, 1950. Pp. 120. \$2.00.
- Educational acceleration, appraisals and basic problems.* Sidney L. Pressey. Columbus: Bureau of Educational Research, Ohio State University, 1949. Pp. 153. \$2.50, paper; \$3.00, cloth.
- Problems of infancy and childhood.* Milton J. E. Senn, Editor. New York: Josiah Macy, Jr. Foundation, 1949. Pp. 156.
- Varieties of delinquent youth.* William H. Sheldon, Emil M. Hartl and Eugene McDermott. New York: Harper and Brothers, 1949. Pp. 899. \$8.00.
- The psychologist in industry.* M. E. Steiner. Springfield, Ill.: Charles C. Thomas, Publisher, 1950. Pp. 107. \$2.00.
- Personnel management for supervisors.* Claude E. Thompson. New York: Prentice-Hall, Inc., 1949. Pp. 192. \$2.95.
- Human relations in modern industry.* R. F. Tredgold. New York: International Universities Press, Inc., 1950. \$2.50.
- Achieving maturity.* Jane Warters. New York: McGraw-Hill Book Co., Inc., 1949. Pp. 349. \$3.00.
- Counseling and discipline.* E. G. Williamson and J. D. Foley. New York: McGraw-Hill Book Co., Inc., 1949. Pp. 387. \$3.75.
- Comprehensive examinations in a program of general education.* Board of Examiners, Michigan State College. East Lansing: Michigan State College Press, 1950. Pp. 165. \$4.00.
- Children absent from school.* New York: Citizens' Committee on Children of New York City, Inc., 1949. Pp. 116. \$1.00.
- Advisory service for students of advertising.* Westport, Conn.: Advisory Service, 1949. Pp. 80. \$1.25.

1949 DIRECTORY

AMERICAN PSYCHOLOGICAL ASSOCIATION

1515 MASSACHUSETTS AVENUE N. W.
WASHINGTON 5, D. C.

In the alphabetical list of 6735 members, the 1949 Directory of the Association gives the names of the members, their addresses, their present positions, their last degrees, and their class of membership. Membership lists for the Divisions of the Association, the lists of Diplomates in the fields of clinical, industrial, and counseling of the American Board of Examiners in Professional Psychology, the By-Laws, and a geographical and institutional index of members are included. The editor is Helen M. Wolfe of the Association staff. 250 pages, \$2.00.

SAMPLE ENTRIES

Humphreys, Lloyd G. School of Education, Stanford Univ, Stanford, Calif. Assoc. prof. educ. and psych. PhD 38. A 5; F 9, 19.

Hunsicker, Mr. Albert L. Committee on Human Development, Univ. of Chicago, Chicago, 37, Ill. Stud. MA 39. A.

Hunt, Dr. Howard F. Dept. Psych, Univ. of Chicago 37, Ill. Assoc. prof. PhD 43. F 12.

Hunt, Dr. J. McV. Institute of Welfare Research, Community Service Society, 105 East 22nd St, New York 10, N. Y. Dir. PhD 33. Dipl.-Cl. F 3, 8, 9, 12.

Hunt, Mary Louise 1252 Talbert St. S. E, Washington, D. C. A '48.

Hunt, Dr. Thelma Dept. Psych, George Washington Univ, Washington 6, D. C. F 5, 12.

Hunt, Prof. William A. Dept. Psych, Northwestern Univ, Evanston, Ill. Prof. psych. and biol. PhD 31. Dipl.-Cl. F 2, 3, 8, 9, 12, 19, 20.

Hunt, Mr. Wilson L. Boston State Hosp, 591 Morton St, Dorchester Center 24, Mass. Clin. psych't. AM 47. A '49.

Hunter, Dr. Elwood C. Dept. Education, Tulane Univ, New Orleans 15, La. Head of dept. PhD 35. A 5, 15.

Journal of Applied Psychology

VOL. 34, No. 3

JUNE, 1950

The Administrative Judgment Test *

Milton M. Mandell

United States Civil Service Commission, Washington, D. C.

The development of valid methods for selecting persons to carry the primary responsibility of directing our large corporations and government agencies is still a major task in the field of personnel administration.

Ample justification for extensive study of this question is found in the fact that analysis of unsuccessful organizations so often indicates that the lack of success is due to lack of administrative ability.

Further inducement may be found in the comment of the late Professor John G. Jenkins that psychologists should be willing to undertake projects in areas of great significance even though the experimental conditions are not so precise as would be desirable.

Some psychologists have recognized the qualitative importance of the field of administrative selection. The broad-gauge work of Carroll Shartle and his group at Ohio State University in the general field of administration, with some attention to selection, should produce fundamental facts that will be of invaluable aid to those interested in the selection of executives and administrators.

One can venture the opinion that while much remains to be known in this field, there are already some established facts which can be used as working hypotheses. For example, it is reasonable to hypothesize that there are both common elements and special elements in all administrative jobs. All positions which can

properly be designated as executive or administrative have certain common elements; however, there are also such special elements as the amount of verbal ability required, the amount of persuasive skill required, the tempo of operations required, and other factors. Any good selection programs for these positions will take into account these common and special factors.

It will be noted that there is no attempt here to define precisely administrative positions or executive positions. Job analyses indicate that program planning and coordination are essential characteristics of administrative and executive positions. Therefore, the following working definition is offered: An executive job or administrative job is one in which more than 50 per cent of the time is devoted to program planning and coordination. This does not, obviously, include all the elements of the administrative job but it does include two elements which will probably always be found.

The purpose of this brief report is to describe the administrative-judgment test which has been developed by the United States Civil Service Commission as part of its program of research in the field of administrative selection. The reason for describing this one test is that it is the test which has given the most consistently satisfactory results and the one on which the most experimentation has been done. While the data below are based on only 171 cases, they represent four different samples and two different types of criteria. The relative consistency of the results among these different samples and the use of different criteria offer a basis for the belief that the test is measuring elements which are essential to administrative success.

The Test. The administrative-judgment test

* Other reports on the experimental work of the United States Civil Service Commission in the administrative field are available in the following articles: Mandell, M. M., The selection and promotion of a personnel staff, *Personnel*, 1948, 25, 125-127; Mandell, M. M., and Adkins, Dorothy C., The validity of written tests for the selection of administrative personnel, *Educ. psychol. Measmt.*, 1946, 6, 293-313; Mandell, M. M., Testing for administrative and supervisory positions, *Publ. Personnel Rev.*, 1948, 9, 190-193.

is in 5-choice form. In the housing studies referred to below, 100 items were used; in the two other studies, 80 items were used. The test attempts to measure broad understanding of the processes of administration. It attempts to measure the understanding of the administrative problems of large organizations, whether government or private. The questions attempt to measure the common elements in the administrative process. They include problems in the relationships between the headquarters and field offices in an organization, and those between research and operating personnel. They also include problems on the timing of programs and the organization of the office of an administrator. The test does not attempt to measure technical knowledge in such fields as personnel or budgeting or accounting. It attempts, as far as possible, to divorce its contents from complete dependence on basic training, and attempts to emphasize problems which can be evaluated on the basis of observation and experience or training. The split-half reliability of the administrative-judgment test is .94 based on the group of 258 cases on which scores were available. Below is a sample item from this test:

Which one of the following administrative situations or problems will most probably occur when direct relations are permitted between a staff specialist employed by the national office of an organization and the operating officials employed in the field offices?

- A) decrease in the feeling of responsibility of national office specialists for the operations of state programs in their specialties
- B) inadequate technical supervision of field office operations
- C) inadequate knowledge in the national office of the competence and qualifications of field office personnel
- D) difficulty in keeping the relations on an advisory basis
- E) subordination of professional considerations to general administrative responsibilities

Criteria Used. Two basic types of criteria have been used in the experiments with this test. The first type, job performance, represents the collective ratings of colleagues and superiors. In all these cases except one, the job performance ratings are a composite of

graphic ratings and paired comparisons; in rating one group of 20 line administrators, only graphic ratings were used. In addition to job-performance ratings, the position grade or salary of the subject was also used as a criterion. An average of more than four independent ratings for each subject in the study was obtained. Because of the small samples involved, practically no cases have been eliminated from the samples because of disagreement among raters. No cases were eliminated from the housing and Veterans Administration studies for this reason. In the case of the Navy study, approximately half of the cases were eliminated because of substantial disagreement among the raters. This greater divergence in the Navy Department as compared with the other groups included in the study can possibly be explained on the basis of the greater size of the Navy Department and the greater number of bureaus that the subjects worked in. The fact that 6 of the 7 correlations represent no elimination of cases lends further weight to the data that are presented. In addition, the scoring key used was determined in advance; in other words, the key that was used is not based upon item analysis on a particular group. In all cases Pearson product-moment correlations were used.

Population. The first group studied were 63 persons in personnel, budget, and organization analysis work in two major housing agencies in the Federal government. They are receiving salaries of between \$3,000 and \$9,400. The group at the Veterans Administration consisted of 42 persons in the personnel office of that agency receiving salaries between \$3,000 and \$6,000. These persons were engaged in all types of personnel work. The Navy group represents persons in several bureaus of the Navy Department engaged in various phases of personnel, office methods, and organization analysis work with salaries between \$3,000 and \$8,400. The last group, also from the housing agencies, consisted of 20 persons, with salaries of between \$7,500 and \$10,000, who are responsible for administering major segments of the Government's housing program.

In addition to the data obtained for the administrative-judgment test, validity coefficients are presented for tests in the general

field of mental ability. For the housing groups, the test involved was the American Council on Education Psychological examination for college freshmen. The total score on the test was used as the predictor. For the other groups, the test consisted of 25 vocabulary items in multiple-choice form. The reliability of the vocabulary test is .84 for a sample of 42 cases with a standard deviation of 5.6, using Kuder-Richardson Formula No. 21. The purpose in presenting these data for mental-ability tests is to demonstrate that the relative validities for this test are substantially lower, in general, than for the administrative-judgment test, despite the fact that the intercorrelation between the administrative-judgment test and the American Council on Education test is $+.69$ while the intercorrelation between the administrative-judgment test and the vocabulary test is $+.59$. A hasty inspection of these intercorrelations would lead to the belief that the mental-ability and administra-

tive-judgment tests are measuring the same factors; actually, the validity coefficients belie this conclusion. The data would indicate that the only proper conclusion is that it is not necessary to use a mental-ability test along with the administrative-judgment test because the multiple correlation of these two tests would not be sufficiently greater to justify the addition of the mental-ability test.

Additional Test Group. The latest trial of the administrative judgment test involved its administration to 30 persons being trained for line and staff administrative positions in the State Department. These persons receive salaries of approximately \$3,000 a year. The criterion was the collective opinion of the supervisors for whom they have worked during their period of internship. The Pearson product-moment correlation for the administrative judgment test with this criterion was $+.60$. The validity coefficient, using the same criterion, for the vocabulary test referred to above was $+.23$.

Table 1

Product-Moment Correlations for the Administrative-Judgment and Mental-Ability Test with Job Performance and Grade Criteria

Agency	N	Criterion	Test of Administrative Judgment	Test of Mental Ability
1. Housing	63	Job Performance	$+.49$	$+.30$
	63	Grade	$+.56$	$+.38$
2. VA	42	Job Performance	$+.50$	$+.52$
	42	Grade	$+.52$	$+.26$
3. Navy	22	Job Performance	$+.51$	$+.13$
	46	Grade	$+.28$	$+.21$
4. Housing	20	Job Performance	$+.68$	$+.64$

Summary

The median validity coefficient for the administrative-judgment test is $+.51$; the median validity coefficient for the mental-ability test is $+.30$. Six of the seven coefficients for the administrative-judgment test are significant at the 1 per cent or 5 per cent level of confidence while the significance level of the coefficients for the mental-ability tests are in general much lower.

These data are offered as a basis for further experimentation in other situations in order to determine the value of the administrative-judgment type of test for executive positions.

Received September 9, 1949.

Menstruation and Industrial Efficiency. II. Quality and Quantity of Production

Anthony J. Smith

University of Kansas

The present paper continues the report of an investigation of the relationships between the various phases of the menstrual cycle and certain measures of industrial efficiency. In all, four criteria of industrial efficiency were studied. In a previous paper (5) the relations between menstrual function and absence rate and activity level were examined.

In view of the common contention that accuracy of performance deteriorates during the premenstrual and menstrual phases, it would seem to be desirable to investigate this hypothesis in an industrial situation. To the knowledge of the author, such a study had not been undertaken previously.

Furthermore, it is frequently assumed that quantity of production decreases during certain phases of the menstrual cycle. Several reports of experimental studies are available on this point, but the results are not consistent. Kirihaara (3) and Meeker (7) report detrimental effects related to the menstrual phase, whereas Nowikowa (2), Gorkin and Brandis (2), and Anderson (1) report no differential effects. The area is obviously in need of further investigation.

Procedure and Analytic Techniques

Aircraft Factory. Twenty-nine women employed in the electrical department of an aircraft factory were studied over a period of forty-one days.¹ All of their work underwent a routine inspection and complete records were kept including the employee's clock number, the unit number, the date, and the defects discovered. In order to obtain data on menstrual function, these women were assembled on company time by the personnel department and requested to participate in an experiment in which the lengths of the menstrual phase and cycle were being studied. All of the

¹ The women who served as subjects in this and the following factories were, with the exceptions noted, the same women described in the first paper (5).

women agreed to participate and later submitted daily menstrual data to a member of the physical education department.

In analyzing the data, the menstrual cycle was broken down into a five day premenstrual phase, the menstrual phase (period of flow), a seven day postmenstrual phase, and an intermenstrual phase.

The effect of the menstrual cycle upon quality of production was determined by computing the number of error days (during which at least one piece was rejected) and the number of errorless days for all women for each of the four phases of the cycle. The data were then recorded in a two-by-four table and the significance of the variations tested by means of the chi-square test. The simpler menstrual-non-menstrual analysis was also performed.

An attempt was made to derive a measure of quality of production that was more sensitive than that involving the mere presence or absence of defective units. However, after several conferences with the supervisory staff it was apparent that there could be no agreement that would permit one to locate on a quality continuum inspection records differing with respect to kinds and numbers of errors. Hence, the cruder measure was retained.

Parachute Factory. The final criterion to be investigated was quantity of production. A group of employees in a parachute factory was contacted by the company nurse and each woman was given an "explanation" of the experiment similar to that previously described. Forty-six women offered to participate and supplied the nurse with the necessary data on menstrual function. The women in this company were paid on a piece rate system with the result that the personnel department maintained accurate individual records including daily production, kind of work, and number of hours worked.

For various reasons production records were not available for five of these subjects. Con-

sequently, the analysis was based on data from the remaining forty-one women.

Examination of individual records disclosed that the number of hours worked per day varied from person to person, from shift to shift, and also for the same individual. This made it necessary to determine for each day the average hourly rate of production for each individual performing her usual work.

It became evident that differences in efficiency among the various women and differences in the types of units produced on the diverse jobs would make the analysis of these raw scores (hourly production rate) undesirable, because variations at the many ability levels and among jobs would then have unequal effects. To meet this difficulty comparable derived scores had to be computed.

At this point in the analysis, each individual was treated separately. The individual's average hourly rates through the total interval investigated were examined. A measure of variability was obtained (standard deviation) and each single hourly rate was replaced by a standard score. Once these standard scores had been derived, individual consideration was discontinued.

The first analysis involved those forty-one subjects for whom production records were available. A distribution of all standard scores falling within the first intermenstrual

period was made and the mean and standard deviation were computed. Similar distributions were made for the subsequent phases (premenstrual, menstrual, postmenstrual, intermenstrual, etc.). These means were then graphed with average standard scores for each phase being plotted along the ordinate while time measurements were recorded along the abscissa. (As an example, see Figure 1.)

The matter of plotting the various elements of the cycles along the time axis posed a further problem. These phases could not be plotted at equal intervals along the axis for they did not represent equal periods of time.

Consequently, each phase was represented by a distance that was in the same proportion to the total distance along the axis that the number of days in the given phase was to the total number of days studied. When the mean production in standard units was plotted for each phase of the cycle, the points were entered at the midpoints of their corresponding intervals along the abscissa.

It will also be seen that there is an increase in mean production in succeeding time periods. This would indicate that a learning factor was operating, as might be anticipated. Improvement, whether it be small or great, would be reflected by an analysis involving standard scores. Furthermore, even assuming that some of the subjects have reached a plateau

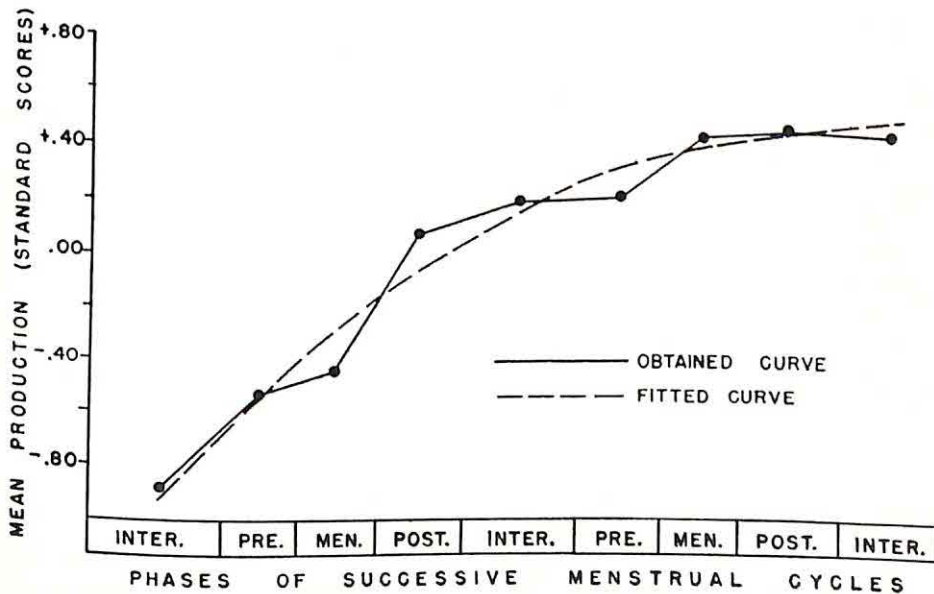


FIG. 1. Mean rates of production during successive menstrual phases. Parachute factory: Ages 29-38.

Menstruation and Industrial Efficiency. II. Quality and Quantity of Production

Anthony J. Smith

University of Kansas

The present paper continues the report of an investigation of the relationships between the various phases of the menstrual cycle and certain measures of industrial efficiency. In all, four criteria of industrial efficiency were studied. In a previous paper (5) the relations between menstrual function and absence rate and activity level were examined.

In view of the common contention that accuracy of performance deteriorates during the premenstrual and menstrual phases, it would seem to be desirable to investigate this hypothesis in an industrial situation. To the knowledge of the author, such a study had not been undertaken previously.

Furthermore, it is frequently assumed that quantity of production decreases during certain phases of the menstrual cycle. Several reports of experimental studies are available on this point, but the results are not consistent. Kirihaara (3) and Meeker (7) report detrimental effects related to the menstrual phase, whereas Nowikowa (2), Gorkin and Brandis (2), and Anderson (1) report no differential effects. The area is obviously in need of further investigation.

Procedure and Analytic Techniques

Aircraft Factory. Twenty-nine women employed in the electrical department of an aircraft factory were studied over a period of forty-one days.¹ All of their work underwent a routine inspection and complete records were kept including the employee's clock number, the unit number, the date, and the defects discovered. In order to obtain data on menstrual function, these women were assembled on company time by the personnel department and requested to participate in an experiment in which the lengths of the menstrual phase and cycle were being studied. All of the

¹ The women who served as subjects in this and the following factories were, with the exceptions noted, the same women described in the first paper (5).

women agreed to participate and later submitted daily menstrual data to a member of the physical education department.

In analyzing the data, the menstrual cycle was broken down into a five day premenstrual phase, the menstrual phase (period of flow), a seven day postmenstrual phase, and an intermenstrual phase.

The effect of the menstrual cycle upon quality of production was determined by computing the number of error days (during which at least one piece was rejected) and the number of errorless days for all women for each of the four phases of the cycle. The data were then recorded in a two-by-four table and the significance of the variations tested by means of the chi-square test. The simpler menstrual-nonmenstrual analysis was also performed.

An attempt was made to derive a measure of quality of production that was more sensitive than that involving the mere presence or absence of defective units. However, after several conferences with the supervisory staff it was apparent that there could be no agreement that would permit one to locate on a quality continuum inspection records differing with respect to kinds and numbers of errors. Hence, the cruder measure was retained.

Parachute Factory. The final criterion to be investigated was quantity of production. A group of employees in a parachute factory was contacted by the company nurse and each woman was given an "explanation" of the experiment similar to that previously described. Forty-six women offered to participate and supplied the nurse with the necessary data on menstrual function. The women in this company were paid on a piece rate system with the result that the personnel department maintained accurate individual records including daily production, kind of work, and number of hours worked.

For various reasons production records were not available for five of these subjects. Con-

sequently, the analysis was based on data from the remaining forty-one women.

Examination of individual records disclosed that the number of hours worked per day varied from person to person, from shift to shift, and also for the same individual. This made it necessary to determine for each day the average hourly rate of production for each individual performing her usual work.

It became evident that differences in efficiency among the various women and differences in the types of units produced on the diverse jobs would make the analysis of these raw scores (hourly production rate) undesirable, because variations at the many ability levels and among jobs would then have unequal effects. To meet this difficulty comparable derived scores had to be computed.

At this point in the analysis, each individual was treated separately. The individual's average hourly rates through the total interval investigated were examined. A measure of variability was obtained (standard deviation) and each single hourly rate was replaced by a standard score. Once these standard scores had been derived, individual consideration was discontinued.

The first analysis involved those forty-one subjects for whom production records were available. A distribution of all standard scores falling within the first intermenstrual

period was made and the mean and standard deviation were computed. Similar distributions were made for the subsequent phases (premenstrual, menstrual, postmenstrual, intermenstrual, etc.). These means were then graphed with average standard scores for each phase being plotted along the ordinate while time measurements were recorded along the abscissa. (As an example, see Figure 1.)

The matter of plotting the various elements of the cycles along the time axis posed a further problem. These phases could not be plotted at equal intervals along the axis for they did not represent equal periods of time.

Consequently, each phase was represented by a distance that was in the same proportion to the total distance along the axis that the number of days in the given phase was to the total number of days studied. When the mean production in standard units was plotted for each phase of the cycle, the points were entered at the midpoints of their corresponding intervals along the abscissa.

It will also be seen that there is an increase in mean production in succeeding time periods. This would indicate that a learning factor was operating, as might be anticipated. Improvement, whether it be small or great, would be reflected by an analysis involving standard scores. Furthermore, even assuming that some of the subjects have reached a plateau

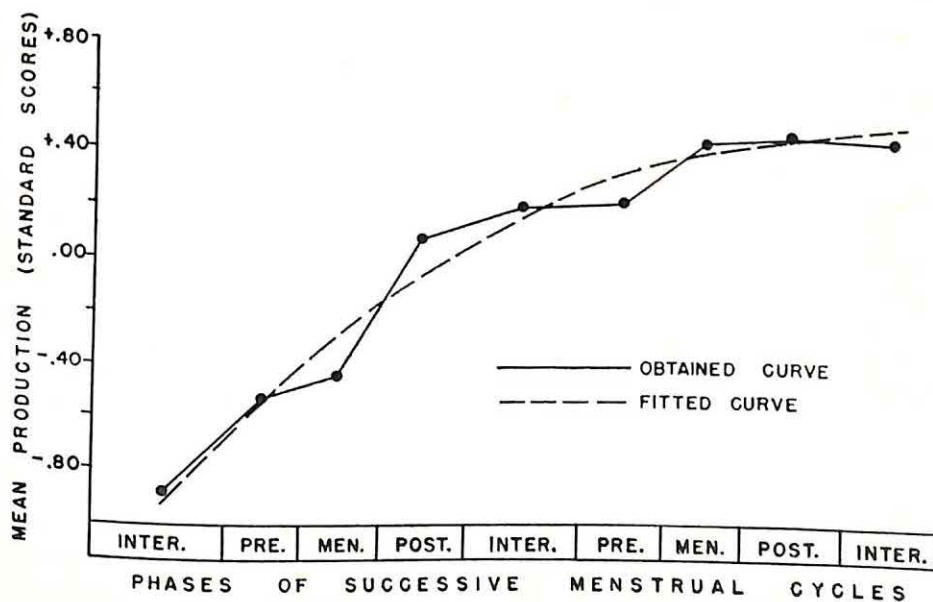


FIG. 1. Mean rates of production during successive menstrual phases. Parachute factory: Ages 29-38.

while the remainder are improving to some extent, the composite curve would exhibit this learning.

The learning factor had to be eliminated before other variations in performance could be studied. Ideally, this would be effected by fitting a theoretical learning curve to the data and, thereafter, considering only variations about the curve. However, in this case no theoretical learning curve was available and it was necessary to fit a smooth curve to the data that would make the deviations about it a minimum. The curves throughout this study were fitted by inspection.

From this point on, the fitted curve was treated as being equivalent to a zero line and the positive and negative deviations of the points about it were computed. These values were the new mean production measures independent of learning, expressed in terms of standard scores.

This was followed by a combination from similar periods of these new scores. A composite intermenstrual mean was derived by combining the means from the three intermenstrual periods studied. Similar composite means were obtained for the other components of the menstrual cycle. Following this, composite variances for the four phases were computed and were then tested for homogeneity (4). With these four composite means and four composite variances available, it was possible to make analyses of the differential effects of the various parts of the menstrual cycle by means of an analysis of variance technique for each situation in which the condition of homogeneity of variances was realized. In each instance, the statistic used was Snedecor's *F* (6).

This general approach was utilized in each of the thirteen remaining analyses of variations in production. The thirteen conditions analyzed include those to be found in Table 2, as well as the first shift and average mental difficulty groups.

Garment Factory. Quantity of production was also investigated at a local garment factory. Data on rate of production were collected by the business manager of the union, who requested that the employees provide her with records of their earnings, presumably for use by the union in subsequent discussions with

management. This approach could conceivably have had some effect on earning rate through the factor of suggestion. However, there was no reason to expect a differential menstrual effect. Special forms were prepared covering weekly periods but requesting daily information on number of hours worked, dozens of units completed, and ticket (daily) earnings.

At approximately the same time, permission was granted by the union to approach individual members and solicit their participation in a research project. Immediately thereafter, a small group of sixteen women was selected to act as potential subjects. These women were contacted by the author's wife during their lunch hour and informed of the apparent "purpose" of the experiment. They then agreed to provide the information requested.

It was decided at a conference involving the personnel manager, the union business agent, and the author that daily earnings were the best available measures of efficiency. The number of units completed was unsatisfactory because the women worked at different jobs requiring varying amounts of time for different units. Earning rates for each of the types of work had been determined by joint labor-management study and consequently reflected ability more adequately. These employees worked on a piece rate basis. Their work was inspected and they were required to reject all rejected units. As a result, daily earnings were measures of quality and quantity of production.

The analysis of the variations in earning rate was carried out in a manner similar to that employed at the parachute factory. Average hourly earning rates were computed and these were transformed to standard scores (each in terms of the subject's own performance). Distributions of the standard scores of all subjects during each of the phases of the menstrual cycle were obtained. The mean scores for each phase were then determined and plotted. A curve was fitted to these points to eliminate the effects of learning. The deviations of the points from this curve were computed and were treated as the new mean scores. Means for similar phases were combined, as were the variances and the variances were then tested for homogeneity. Finally, an analysis

Table 1

Quality of Production: Frequency of Occurrence of Error and Non-Error Days During the Phases of the Menstrual Cycle

	Premenstrual	Menstrual	Postmenstrual	Intermenstrual
Error Day	18 (18.78)*	17 (20.16)	19 (24.41)	48 (38.69)
Non-error Day	132 (131.22)	144 (140.84)	176 (170.59)	261 (270.31)

* Numbers in parentheses refer to theoretical frequencies.

of variance was performed on the derived composite means and composite variances.

Results and Discussion

Quality of Production. The analysis of the four phase data obtained from inspection records (see Table 1) revealed differences that were not significant (P equals .20). The gross analysis contrasting menstrual with nonmenstrual days yielded much higher agreement between theoretical and obtained frequencies (P equals .60). The trend was in the direction of fewer error days during the menstrual period.

It was felt that some of the women might manifest an obvious detrimental effect during the menstrual phase and that such an effect would be obscured by group treatment. Although individual analyses could not be undertaken satisfactorily, an examination of individual records disclosed no tendency toward a greater proportion of error days in the menstrual period than during the nonmenstrual interval. Those persons who performed poorly during the period of flow tended to perform poorly at all times.

The failure to discover a significant decrease in the quality of production during the menstrual phase might conceivably be explained on the grounds that accuracy was maintained at the expense of lowered production. A consideration of the results derived from the parachute and garment factories would seem to render this interpretation untenable.

Quantity of Production. In checking the homogeneity of the variances of production rates during the several phases of the menstrual cycle, significant differences were encountered in only two instances. On the first shift, while the differences were significant, variability

was highest during the postmenstrual period and lowest during the premenstrual period. Women working on jobs described as being of average mental difficulty also revealed significant differences in variability, with variability being greatest during the intermenstrual phase and lowest during the premenstrual phase. This decreased variability was not accompanied by a decrease in production. Obviously, no detrimental menstrual effect is indicated.

In the analysis of the production rates of the remaining testable groups, two yielded sig-

Table 2

Rate of Production: Values of F and Corresponding Probability Values

	Number of Days Analyzed	F	Probability
Parachute Factory			
All Subjects	1711	1.67	.19
Shift			
Second	553	1.88	.15
Third	567	1.09*	>.20*
Age			
18-28	438	2.76*	>.20*
29-38	576	1.82*	>.20*
39-50	697	1.18	>.20
Mental Difficulty			
Simple	252	2.24	.10
Difficult	378	3.55	.02
Physical Difficulty			
Average	1431	2.03*	>.20*
Strenuous	280	3.17	.03
Standing	707	1.16	>.20
Sitting	1004	1.18	>.20
Garment Factory			
All Subjects	457	4.66*	.15*

* Variance among groups is smaller than variance within groups.

nificant differences (see Table 2). Women engaged in jobs involving a relatively high level of mental difficulty manifested significant differences in production rate. However, their lowest production occurred during the premenstrual phase, with highest production appearing in the menstrual phase and being of such a level as to counteract the premenstrual loss. The second group of women working on jobs demanding strenuous physical activity displayed high postmenstrual production and low intermenstrual production.

In brief, variability in production on the jobs studied in this investigation does not increase during the premenstrual or menstrual phases as has been claimed. If anything, it would appear to decrease. Furthermore, rate of production does not seem to decrease during the period of flow or during a period of possible "premenstrual tension," with the exception of the women performing work of a high level of mental difficulty. As a matter of fact, there is evidence that a drop in production occurs in the intermenstrum in one instance. It may be that both of these "significant" differences could have occurred by chance. However, in the groups not yielding significant differences the intermenstrual period is often characterized by low production, whereas the postmenstrual period is usually one of high production.

Summary

This second part of the investigation was undertaken to determine the effects of the various phases of the menstrual cycle upon industrial efficiency as reflected in quality and rate of production.

A total of eighty-six women in the aircraft and garment industries served as subjects in the three component studies. Approximately thirty-eight hundred individual working days were ultimately analyzed in the complete study.

In each possible instance, two analyses were performed. In the gross analysis, menstrual performance was contrasted with nonmenstrual performance. The more intensive analysis was concerned with a comparison of performances during the premenstrual, menstrual, postmenstrual and intermenstrual phases.

In the study of the employees of the parachute factory, the subjects besides being examined as a group were classified into subgroups according to shift, mental difficulty of work, physical difficulty of work, age, and standing vs. sitting jobs to determine possible differential effects of the cycle.

The study of quality of production, measured in terms of the presence or absence of error days (days on which defective units were worked), shows that variations are generally small and unrelated to the component phases of the menstrual cycle.

Significant differences in variability of production rate occur in two of the subgroups. In one case, greatest variability is found in the postmenstrual period, while in the other, it occurs in the intermenstrual period.

All of the analyses of production rate except two reveal no statistically significant differences among the phases of the menstrual cycle. In one group intermenstrual production is low, while the second group displays low premenstrual production which is offset by high menstrual production.

No one phase of the menstrual cycle yields losses in efficiency more frequently than any other phase. Where significant differences occur they would appear to be the result of situational determinants rather than menstrual function.

Received September 20, 1949.

References

1. Anderson, Mary. Some health aspects of putting women to work in war industries. In *Seventh annual meeting, Industrial Hygiene Foundation of America, Inc.*, Nov. 10-11, 1942, 165-169.
2. Gorkin, S., and Brandis, S. Einfluss der Menstruation auf einige psychophysiologische Funktionen und auf Arbeitsfähigkeit der Frau. *Arbeitsphysiologie*, 1936, 9, No. 3.
3. Kirihaara, H. Functional periodicity. *Rep. Inst. Sci. of Labour*, Kurasiki, Japan, 1932, No. 14.
4. Rider, P. R. *An introduction to modern statistical methods*. New York: John Wiley, 1939.
5. Smith, A. J. Menstruation and industrial efficiency: I. Absenteeism and activity level. *J. appl. Psychol.*, 1950, 34, 1-5.
6. Snedecor, G. W. *Statistical methods*. Ames: College Press, Inc., 1938.
7. Women's Work (Anon.). Occupation and Health. Int. Lab. Off., Geneva, 1930-1934, No. 152.

Cross-Validation of Clerical Aptitude Tests

Edward N. Hay

Aptitude Test Service, Swarthmore, Pa.

Tests have been used for fifteen years in the selection of clerical workers at the Pennsylvania Company for Banking and Trusts, Philadelphia. In 1941 a study was made of the prediction of success in machine bookkeeping,¹ using speed of posting as the criterion. A battery of three tests, Number Series Completion, Name Finding, and Minnesota Numbers, was found to give excellent results in predicting success in bookkeeping.

The present study was undertaken to determine whether it was possible to predict success of key-punch operators with the same tests. Since there appeared to be no objective way of ranking key-punch operators according to production because of differences in their jobs, their performance was rated by their supervisors. The ratings were under the general title of "performance," and were divided into six levels, with appropriate headings, as follows: Group I. Quantity and quality of production are above normal, shows initiative, capable of handling greater responsibilities; Groups II and III. Quantity and quality of production are entirely satisfactory, shows some ability in handling unusual problems, learns new work with average instructions. (This section was divided into two sub-sections.); Groups IV and V. Quantity and quality of production are just satisfactory, requires more supervision than the average employee. (This group was subdivided into two.); and Group VI. Quantity and quality of production unsatisfactory.

In rating the clerks each one was compared with the others as well as rated on the characteristics indicated by the headings just given.

The ratings of the key-punch operators were made by the tabulating department manager who supervised clerks engaged in all of the operations usual to Remington-Rand punch card tabulating equipment. These particular clerks were employed without experience and were then trained to operate the Remington-

Rand key-punch machine, which has a typewriter keyboard for letters and a special bank of keys for the numbers. Most of these operators were under 20 years of age and nearly all were under 25. Few had had office experience or typing training.

The Criterion

Eighty-two key-punch operators were rated who had been hired from 1941 to 1944, had remained on the job long enough to become proficient and to establish records which could be rated, and had taken the three tests referred to above. The number of key-punch operators rated in each of the six groups was as follows: Group I. 6; Group II. 16; Group III. 31; Group IV. 13; Group V. 9; and Group VI. 7.

The supervisor who did the rating was aided by several assistants who were group leaders. The range of test scores was somewhat greater than in normal times because the labor market was tight in these years, and it was frequently necessary to lower the hiring requirements.

The tests used in this study were: (a) Minnesota Clerical—Number Checking;² (b) Number Series Completion;³ and (c) Name Finding.³

The Minnesota Clerical Test needs no description here. The Number Series Completion test was the form used by Guilford in the Nebraska Revision of Alpha and was drawn from several Alpha series. The Name Finding test is modeled on the number test of IER Clerical. It consists of 25 names on the front of a sheet; for example, Allen B. Smith. On the back of the sheet are groups of four names together: A. C. Smith, A. B. Smyth, A. B. Smythe, and A. B. Smith. The subject reads the name on the front of the sheet and then turns the sheet over in order to check the correct choice. This operation is very much like that performed by a bookkeeper in turning

² Obtainable from The Psychological Corporation, 522 Fifth Ave., New York.

³ Obtainable from Aptitude Test Service, Swarthmore, Pa.

¹ Hay, E. N. Predicting success in machine bookkeeping. *J. appl. Psychol.*, 1943, 27, 483-493.

Table 3

Scatter Diagram Showing Relation Between Ratings of Key Punch Operators and Scores on the Name Finding Test

Rating Groups	Scores on Name Finding Test																	Total
	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21		
I. Good								1			1	3	1				6	
II. Good								2	4	4	2		2	1	1		16	
III. Good						1	4	7	6	6	4	2	1				31	
IV. Poor			1		1	3	3	2			1		1	1			13	
V. Poor	1		1		1		1		2	2						1	9	
VI. Poor									1	5			1				7	
Number "good"	0		0		0	1	4	9	11	10	7	5	4	1	1	0		
Number "poor"	1		2		2	3	4	2	3	7	1	0	2	1	0	1		
Cumulative "good"	53		53		53	53	52	48	39	28	18	11	6	2	1	0		
Cumulative "poor"	29		28		26	24	21	17	15	12	5	4	4	2	1	1		
Cumulative total	82		81		79	77	73	65	54	40	23	15	10	4	2	1		
% Pass	100		99		96	94	89	79	66	49	28	18	12	5	2	1		
% "Good" of those who pass	65		65		67	69	71	74	72	70	78	73	60	50	50			
			Failure 1/8				4/4 Doubtful	Success 48/17 or 2.8/1										

the first three rating groups were the "good" clerks and those in the last three groups were the "poor" clerks. There are 53 of the former and 29 of the latter. Of the clerks who scored less than 130 on Minnesota Numbers, there are about as many poorer ones as better ones. Accordingly, the scores below 130 on this test fall in "uncertain" territory. Above that point the ratio of better to poorer is $4\frac{1}{2}$ to 1.

In the case of the Name Finding test, as shown in Table 3, the ratio of better to poorer (who score above 12) is not so high — 2.8 to 1. However, except for "uncertainty" with a score of 12, the prediction from scores of 11 and less is strongly toward "failure," or poorer clerks, the ratio being 8 to 1 for failure. Another difference between the tests is seen in the critical percentages. The "passing" group with Minnesota Numbers (making 130 or more) comprises only 40% of the criterion group, of which 82% are good and 18% poor, whereas 79% of the criterion group exceeded the critical score on Name Finding and 74% of them were the higher rated clerks against 26% lower rated. The coefficients of correlation with the

rating criterion are .302 and .260, respectively, for Minnesota Numbers and Name Finding. These examples show how easy it is to find the best cutting scores for single tests even though the correlations are low.

Multiple Cutting Scores

A study of the scattergrams suggested that a combination of critical scores from two or three tests might be even more effective in discriminating between "good" and "poor" key-punch operators than the cutting score from just one test. The object was to find a combination of scores that were passed by a large percentage of the clerks rated "good," and that will eliminate as many as possible of those rated "poor." In order to do this, it was decided to list the scores for individuals by rating groups. Table 4 shows the scores of the applicants who were rated in the second and fifth of the six rating groups. The rest of the list is omitted to save space.

Table 5 shows the results of trying different combinations of cutting scores. The combina-

tions listed here are but a fraction of the many trials that would theoretically be necessary to exhaust all the possibilities of combinations of cutting scores. But from a study of the list (given in part in Table 4) and the scattergrams, these seemed to be the only combinations that offered any promise. In Table 5 the critical score combinations are arranged in descending order, assuming that hiring standards would

be lowered successively as the labor market tightened. Combinations of test scores for this group give better results at all score levels than the multiple regression equation possibly because the distributions are not linear. The best combinations for different levels of hiring are shown in Table 6.

In studying the list of scores, it was found that there were several individuals in Group

Table 4

Scores on Tests Taken by Key Punch Operators at Time of Application for Employment

Note: Scores are given for individuals in two of the six rating groups. The scores listed at the top of the 5 columns at the right are some of the combinations of cutting scores on indicated tests which were selected for trial. An X was placed after the scores of each individual who failed to "pass" the particular combination of cutting scores. By counting these X's in a given column, it is possible to show the number of good and poor clerks who would have been eliminated if hiring had been based on these standards. These 5 combinations are representative of the different combinations which were tried.

	(1) Minnesota Numbers	110	110	110	—	110
	(2) Number Series	9	8	7	9	9
	(3) Name Finding	13	12	11	13	—
	(1)	(2)	(3)			
Group II						
No. 1	136	13	14			
No. 2	150	15	18			
No. 3	116	17	15			
No. 4	97	11	13	X	X	X
No. 5	141	9	16			
No. 6	110	10	13			
No. 7	141	12	15			
No. 8	126	11	18			
No. 9	186	11	19			
No. 10	126	11	16			
No. 11	139	14	15			
No. 12	125	14	14			
No. 13	142	14	14			
No. 14	114	11	14			
No. 15	110	12	15			
No. 16	142	13	20			
Group V						
No. 1	146	7	15	X	X	X
No. 2	118	8	14	X		X
No. 3	118	8	14	X		X
No. 4	112	10	15			
No. 5	140	8	10	X	X	X
No. 6	107	10	12	X	X	X
No. 7	125	7	6	X	X	
No. 8	117	12	21			X
No. 9	127	7	8	X	X	15
Number passed in Group II				15	15	2
Number passed in Group V				2	4	5

Table 5

Possible Selections of Key Punch Operators from Test Scores, Using Indicated Cutting Scores on the Given Tests and Showing the Per Cent Who Pass and the Per Cent Rated "Good" of Those Who Pass.

Note: N is 82, with 53 rated "good" and 29 rated "poor".

Minn. Nos.	No. Series	Name Finding	Per Cent Who Pass	Per Cent Rated "Good" of Those Who Pass
110	9	13	59	86
110	8	12	73	78
110	9		70	79
110		13	70	79
	9	13	68	79
	9	12	78	76
		13	79	74
110	7	11	77	75
		12	89	71
110			85	70
	8	12	84	72
	9		83	71

VI who made high scores on all three tests. Some of these were found to be "problem cases." Of one it was said, "Hated the work; was transferred to credit analysis." Of another, "Hired as an experiment; had only one arm. Left because of a nervous condition." Of still another, "Had ability but refused to cooperate. Poor health and attitude." These individuals were, of course, left in the sample.

These comments help to explain, however, why it is never possible to get 100 percent success in selecting key-punch operators, or any other workers, on the basis of test results alone.

Comparing Results with Previous Study

It is interesting to compare the results of the present study with the earlier study of 39 bookkeepers in the same organization. Since the same clerical aptitude tests were used, the value of using multiple cutting scores can be studied in this sample also. In the bookkeeper group, slightly better prediction was obtained by means of the multiple regression formula than with multiple cutting scores. The R in this case was .70 as found by the Doolittle method, the tests which contributed being: (1) Minnesota Numbers, (2) Number Series Completion, and (3) Name Finding. The intercorrelations are given in Table 1. Here the criterion was actual rate of production, usually a more reliable criterion than ratings of supervisors. An examination of the results secured with multiple cutting scores in the bookkeeper group confirms the satisfactory prediction attained in the key punch group. Table 6 lists the successful predictions from the same critical scores shown in Table 4 when applied to the 39 bookkeepers. The dividing line between "good" and "poor" clerks was taken at the

Table 6

Prediction Results for the Key Punch Group and the Bookkeeper Group Using Indicated Cutting Scores on the Three Tests

Note: 82 Key Punch Operators, 53 rated "good" and 29 rated "poor"; 39 Bookkeepers, 29 "good" and 10 "poor" producers

Cutting Scores			Key Punch Group		Bookkeeper Group	
Minn. Nos.	Number Series	Name Finding	Per Cent Who Pass	Per Cent Rated "Good" of Those Who Pass ¹	Per Cent Who Pass	Per Cent Rated "Good" of Those Who Pass
110	9	13	59	86	54	86
110	—	13	70	79	79	81
110	9	—	70	79	61	83
—	9	13	68	79	56	86
—	—	13	79	74	84	79
—	9	—	83	72	64	84
110	—	—	81	71	95	73

production rate of 105, which was the production average of the group at the time the study began in 1937; 29 operators achieved 105 or better and 10 operators achieved less than 105. In Tables 1 and 6 the results of both studies are given for comparison and to show how one study confirms the other.

In view of the way in which the results of the bookkeeper group confirm those of the key-punch operators it is interesting to note that the subjects were quite different in several respects. Although the key punch operators were mostly inexperienced at the time of hiring, the bookkeepers were almost all experienced clerks. Their average length of service was 9 years and 2 months. The bookkeeper study took place in the years 1937-1940 and the key punch study after 1944. One group operated the Burroughs adding-bookkeeping machine, and the other was trained to use the Remington-Rand key-punch machine.

Comparison of the two prediction formulae shows the differences in the proportionate contributions of the three tests:

Bookkeepers:

$$\bar{X}_0 = .19 \times \text{Minn. Nos.} + 1.34 \times \text{No. Series} + 1.27 \times \text{Name Finding} + 52.4.$$

Key Punch operators:

$$\bar{X}_0 = .15 \times \text{Minn. Nos.} + .97 \times \text{No. Series} + .35 \times \text{Name Finding} + 67.3.$$

Summary

A study was made of prediction of success in a group of 82 key punch operators on the basis of a battery of three clerical tests. Although the Doolittle multiple R was low, it was possible to make satisfactory predictions on the basis of a combination of cutting scores on the three tests.

The results of this study corroborate the findings of the bookkeeper study made in 1941. The tests used were Minnesota Numbers, Number Series Completion, and Name Finding. They require only fourteen minutes of testing time and predict equally well for experienced and inexperienced clerks.

Received April 12, 1950.

Early publication.

A Test Battery for Actuarial Clerks *

Adam Poruben, Jr.

Personnel Division, Metropolitan Life Insurance Company

In June, 1947, the writer was requested to validate a test battery for the selection of Actuarial Clerks in the Metropolitan Life Insurance Company. At this time, a review of the job description of the Actuarial Clerk, in cooperation with an Assistant Actuary, revealed at least three characteristics necessary for the performance of this job, namely, mental alertness, numerical aptitude, and memory. Because of practical considerations, only a small sample of 12 Actuarial Clerks was available at this time. Six tests, designed to measure the above characteristics, were administered to this group. The rank correlations between test scores and overall performance on the job indicated that five of these tests might have some value for the prediction of success on the Actuarial Clerk job. The rank correlations for the individual tests were fairly low but when the raw scores were transformed into standard scores and a composite score was found for each individual, a rank correlation of $+0.71$ was obtained between the composite scores and overall performance on the job. In spite of the small sample, the writer and management decided to use these tests tentatively until it would be possible to validate the tests on a larger sample. The purpose of this paper is to report the results of such a study.

The Sample

The sample consisted of 125 Actuarial Clerks who had taken the five tests used in the original study. The majority of these Actuarial Clerks were called either Computing or Calculating Clerks. All of these employees had at least six months of service on the particular job they were doing at the time of this study. The average length of service on the job for the entire group was 13.5 months with an S.D. of 7.65 months. The average length of service

with the company for the entire group was 9.5 years with an S.D. of 8.10 years.

The Home Office Clerical Jobs of the Metropolitan Life Insurance Company are at the present time classified into 19 levels, ranging from level A to level S. In order to give the reader an idea of the complexity and responsibility of the various jobs represented in the above sample, the job level distribution of these 125 jobs is given. It is as follows: 15 B's, 14 C's, 25 D's, 18 E's, 17 F's, 16 G's, 9 H's, 3 I's, and 8 J's.

Description of the Tests

1. *Otis Self-Administering Test of Mental Ability.* This test is well known and need not be described in detail. The 30-minute time limit was used. The author reports a test-retest reliability of .92.

2. *L.O.-M.A. 4-M Test.* This test is published by the Life Office Management Association and consists of 39 problems: 11 percentages, 10 fractions, 7 decimals, and 11 numerical reasoning problems. It has a time limit of one hour. The corrected odd-even reliability of this test for the 125 sample used in this study was $+0.91$. A test-retest reliability coefficient of $+0.86$ was also obtained for this test on a small sample of 24 company employees.

3. *Ratio-Proportion Test.* This test was constructed by the writer and consists of 11 problems involving the use of ratio and proportion. Three of the problems call for simple interpolation of tables since considerable interpolation of tables is done on most of the Actuarial Clerk jobs represented in this study. It has a time limit of 27 minutes. Its corrected odd-even reliability was found to be $+0.77$ for the 125-case sample of this study.

4. *Logical Memory Test.* This test consists of a short story written in a single paragraph. The various ideas in this story are separated by diagonal lines to facilitate memorization. The subject reads this story for two minutes and

* Grateful acknowledgment is made to Mr. T. A. Crowther, Mr. V. G. Christman, Mr. V. A. Lane, and Mr. H. L. Rhoades for their assistance in this study.

then reproduces from memory as much of the story as he can. The score is the number of ideas correctly reproduced. No estimate of reliability was possible for this test.

5. *Wesman Personnel Classification Test.* Only Part II of this test was used. This consists of twenty numerical problems. It has a time limit of 10 minutes. Its corrected odd-even reliability for the 125-case sample studied was $+.85$. The author reports a reliability coefficient of $.82$ for 174 chain store clerks.

The Criterion

The criterion consisted of ratings obtained on an experimental rating form. This form consists of eight traits: Knowledge of Work, Quality of Work, Quantity of Work, Ability to Learn, Cooperation, Interest in Work, Attendance, and Punctuality. Ratings on the first six traits and an over-all rating based on all traits were used as criteria in this study. The trait ratings are obtained on a five-degree graphic scale. The scales on Knowledge of Work, Quality of Work, and Quantity of Work have a range of 1 to 20; those for Ability to Learn, Cooperation, and Interest a range of 1 to 10; those for Attendance and Punctuality a range of 0 to 5. The over-all rating is obtained by summing the point ratings on all of the eight traits.

This experimental rating form was designed by the Company in 1948 in order to improve the employee-evaluation procedures. It was tried out on 3,275 non-supervisory clerical employees in September 1948 and on 8,876 non-supervisory clerical employees in February 1949. The results were quite satisfactory. The mean and S.D. of the September 1948 over-all ratings was found to be 69.8 and 11.16, respectively; the mean and S.D. of the over-all ratings on the 8,876 sample was found to be 69.5 and 11.26, respectively. Both of these experimental runs were first explained to the Company managers by the Personnel Officer who urged them to give the best and most unbiased ratings since no official records were to be kept and no administrative actions, such as salary adjustments, were to be taken on the basis of these ratings. Also the managers were urged to rate only employees with at least six months experience and to disregard any ratings which they have given any employee at any

time before these experimental runs. Also the managers had no knowledge of the fact that a second experimental run would be made at the time of the first run. Therefore, it is safe to say that the two ratings on the 3,275 employees were made independent of each other.

As was mentioned before, two ratings were available for 3,275 employees. From this group, 2,210 were selected who were still on the same job at the time of the second rating as the first rating. These 2,210 employees, therefore, had at least six months experience on their job at the time of the first rating and at least one year at the time of the second rating. Most of the jobs held by these 2,210 employees are such that the employee can be trained to do the work in a few months. Thus, at the end of six months the employee should have attained enough skill on his or her particular job that fairly accurate evaluation of his or her work is possible. When the two sets of over-all ratings for these 2,210 employees were correlated a coefficient of reliability of $+.69$ was found.

The means and S.D.'s of these two sets of ratings for the 2,210 employees were not statistically significantly different. The first ratings had a mean of 69.69 and an S.D. of 10.97; the second ratings had a mean of 70.35 and an S.D. of 11.08.

For obtaining estimates of reliability of the trait ratings, the ratings of the employees from the Actuarial Division who were rated twice were used. It was found that there were 171 such employees who were rated twice, who were on the same job at the time of the second rating as the first, and who had at least six months experience on their job at the time of the first ratings. The rate-rerate coefficients of reliability were found to be as follows: $+.64$ for quality of work, $+.65$ for quantity of work, and $+.71$ for ability to learn.

Results

In order to see if the tests differentiated between the outstanding and the poor workers, the test scores of the thirty-two (25.6%) employees with the highest over-all ratings were compared with the test scores of the thirty (24.0%) employees with the lowest over-all ratings. The results are shown in Table 1.

Distributions for all variables were drawn

Table 1
Average Test Scores and Ratings for the Best and Poorest Employees

Variable	Upper Group		Lower Group		<i>t</i>
	Mean	S.D.	Mean	S.D.	
Ratio-Proportion	7.7	2.74	5.9	2.40	4.32*
L.O.-M.A. 4-M	23.6	9.38	18.1	7.71	2.47*
Personnel Classification	12.6	3.95	10.7	3.64	2.00**
Otis Mental Ability	54.4	11.57	49.2	11.09	1.77
Logical Memory	46.1	9.22	44.6	10.73	.60
Over-all Ratings	82.9	3.18	58.5	6.84	17.69*

* Significant at the one per cent level.

** Significant at the five per cent level.

and found to be approximately normal. The scatter plots between the variables indicated that the use of the Pearsonian product-moment correlations was permissible. The means, S.D.'s, test intercorrelations and validity coefficients are shown in Table 2.

The correlations between the tests and ratings on Cooperation, Interest in Work, and Knowledge of Work were not significant at the one per cent level and, therefore, are not shown in Table 2.

Two more correlations were found which do not appear in Table 2. One of these was between length of service with company and over-all ratings. It was found to be .12. The other was between over-all ratings and the job level. This was found to be .32.

The Wherry-Doolittle technique was used to combine the tests into batteries. All the test-criterion correlations were corrected for attenuation in the criterion before this technique was applied. The only combination that gave a higher validity coefficient than the

single tests was the combination of the Ratio-Proportion and Wesman Personnel Classification Tests which resulted in a shrunken multiple *R* of .457 and Beta weights of .333 and .159, respectively, with the criterion Ability to Learn. Apparently there is considerable overlap among the tests.

Discussion

It can be seen from Tables 1 and 2 and the results given in the last section that the Ratio-Proportion Test is the best predictor since it correlated significantly with ratings on Quality of Work, Ability to Learn, and with over-all ratings. The L.O.-M.A. 4-M came out about the second best in that it was found to correlate significantly with the ratings on Quality of Work and Ability to Learn. The Wesman Personnel Classification Test also correlated significantly with two of the criteria of job success—Quality of Work and Ability to Learn. The Otis was found to correlate significantly with only one of the criteria, namely,

Table 2
Test Intercorrelations and Validity Coefficients for 125 Actuarial Clerks *

Mean	S.D.	Description of Variable		B	C	D	E	F	G	H	I
11.4	3.78	Personnel Classification	A	.74	.31	.82	.74	.21	.23	.15	.34
52.0	11.57	Otis Test of Mental Ability	B		.56	.78	.77	.17	.20	.14	.32
45.6	9.99	Logical Memory	C			.36	.41	.10	.12	.08	.20
19.9	8.61	L.O.-M.A. 4-M (Arithmetic)	D				.80	.21	.25	.20	.34
6.9	2.70	Ratio-Proportion	E					.26	.30	.21	.38
71.3	9.89	Total Rating	F						.92	.84	.77
14.1	2.57	Quality of Work	G							.76	.73
13.7	2.34	Quantity of Work	H								.59
7.3	1.21	Ability to Learn	I								

* A correlation of .23 is significant at the 1% level.

Ability to Learn. The Logical Memory Test showed no significant correlations with any of the seven criteria used in this study.

In view of the above results, it can be safely concluded that the Ratio-Proportion, the L.O.-M.A. 4-M and the Wesman Personnel Classification Tests are valid for the prediction of success on the Actuarial Clerk job. Although the validity coefficients are not high, it can be stated that only a relatively small percentage of the employees tested, those with the higher test scores, are considered for selection and placement after their other personnel records such as years of service, attendance, former ratings, etc. are reviewed. Under these conditions even relatively small validity coefficients have some predictive value.

It is not surprising that the tests did not show any validity for the prediction of behavior included under the trait Cooperation nor for Interest in Work. The tests were designed to measure aptitudes and not personality characteristics or interest. The results bear this out.

The analysis in connection with the multiple correlation work indicated that there is considerable overlap among the tests. It appears that a two-test combination consisting of the Ratio-Proportion and the Wesman Personnel Classification Tests has as good or better validity than all five tests combined. The Logical Memory Test definitely does not add to the efficiency of prediction. This is probably because of its poor reliability due to subjective scoring. Also it appears that whatever is measured by the Otis is just as well measured by the L.O.-M.A. 4-M, Ratio-Proportion and the Wesman Personnel Classification Tests. This is not too surprising since fairly large portions of these tests, especially the first two, consist of numerical reasoning problems.

Summary

A battery of five tests, designed to measure mental alertness, numerical aptitude and memory, was tried out on a sample of 125 Actuarial Clerks in the Metropolitan Life Insurance Company. All of these employees had at least six months of service on the particular job they were doing at the time of this study, the average length of service on the job for the entire group being 13.5 months with an S.D.

of 7.65 months. The average length of service with the Company for the entire group was 9.5 years with an S.D. of 8.10 years. Their job levels ranged from B to J.

The criterion consisted of ratings on six traits and an over-all rating. The reliability of the over-all ratings was estimated by obtaining a correlation between two ratings on 2,210 employees, the two ratings being five months apart. All of these employees were on the same job during both ratings, had at least six months experience on their job at the time of the first rating and were rated by the same supervisor at both times. This correlation was found to be $+ .69$. Another measure of the stability of the over-all ratings was obtained by finding the means and S.D.'s of the two sets of ratings. The first ratings had a mean of 69.7 and an S.D. of 10.97; the second ratings had a mean of 70.4 and an S.D. of 11.08. The reliability coefficients for the traits on a sample of 171 employees, ranged from $+ .64$ to $+ .71$.

Two of the tests were found to differentiate significantly between the best 25 per cent and the worst 25 per cent of the employees. These were the Ratio-Proportion and the L.O.-M.A. 4-M Tests. The Wesman Personnel Classification Test differentiated significantly between these two groups at the five per cent level.

The Ratio-Proportion Test was found to be the most valid, having significant correlations, at the one per cent level, with three of the seven criteria of job success, namely, Quality of Work, Ability to Learn and over-all ratings. The L.O.-M.A. 4-M was the second best predictor having significant correlations with Quality of Work, and Ability to Learn. The Wesman Personnel Classification Test also correlated significantly with two of the seven criteria, namely, Quality of Work and Ability to Learn.

Great overlap was found to exist among the five tests. On the basis of the multiple correlation results it appears that the combination consisting of the Ratio-Proportion and the Wesman Personnel Classification Tests has as good or better validity than all five tests combined.

*Received March 2, 1950.
Early publication.*

Changes in Subjective Fatigue and Readiness for Work During the Eight-Hour Shift *

John W. Griffith, Willard A. Kerr, Thomas B. Mayo, Jr.

Illinois Institute of Technology

and

John R. Topal

Belden Manufacturing Company

The Present Research

Although remarkable progress has been made in industrial psychology in recent decades, it is an interesting fact that remarkably little research has been reported on the changes which probably occur in the subjective fatigue and readiness for work of personnel from one part of a standard work shift to another. As Ryan (5) has pointed out, most "fatigue" research has reflected an academic preoccupation either with trying to measure objective fatigue or attempting to define fatigue with precision, the latter task being one which Muscio (3) decided as early as 1921 to be practically hopeless. Because of the relative absence of literature on the subjective feeling changes with work, the problem is largely unmentioned in existing textbooks on personnel and industrial psychology.

Production curves for the work day in various factory operations sometimes are presented for their possible relevance (1, 4) to fatigue, but it is admitted generally that many factors other than fatigue, however defined, determine the production curve. Feelings of tiredness do not necessarily change in expected directions with changes in rate of output. However, to date, the scheduling of work and locating of rest pauses in industry have been done largely according to guesswork or with reference to incidental practical considerations independent of worker readiness.

If considerable consistency is found in worker feelings of tiredness and readiness to work in various types of work, it is possible that such feeling curves will be useful in the more intelligent scheduling of work and rest pauses in business and industry.

The present research entertains the hypothesis that employees in representative types of work possess definite attitudes as to when during the work shift they are most ready to work and when they are most tired. Employing the "tear ballot" technique (2) a measuring device was constructed which obtains from the worker his estimate of when in each half of the eight-hour shift he feels most rested and most tired. All replies of 379 employees were anonymous except that eleven ballots were temporarily coded for a crude test-retest reliability check in addition to other internal consistency evidence. Included in the sample were 232 male manual workers (handlers and sorters of light-to-100-pounds materials). 75 foremen in a rawhide factory, and 72 office workers (48 male, 24 female). Foremen were measured at a regular meeting of the Chicago Rawhide Management Club and the other personnel were measured while at work in a distributing organization (manual workers) and in an electronics plant (office workers). The supervisory and office personnel were regular day shift employees but the 232 manual workers began their various shifts in the three-hour period from 3:30 to 6:30 p.m.

Results

Obtained subjective reports of tiredness and readiness for work at various hours of work were analyzed with respect to age, sex, and type of work performed. Repeat-test reliability coefficients for small groups ranged from .69 to .92.

Age. Office and manual workers were

* Senior author is W. A. Kerr.

studied separately as to possible age differences in work feelings. Each group was divided into younger (20-35) and older (36-65) personnel and per cent of each age group feeling tired (and rested) at each of the eight hours of the shift was calculated. Considering each half of the work shift separately, normal chance expectancy, assuming no change in work feelings with successive hours of work, would yield 25 per cent response at each hour of each half of the work shift. Actually, conspicuous changes in both tiredness and readiness for work are reported by both older and younger workers in successive hours of work. Older workers, both office and manual, report greater average feeling deviations from chance expectancy than do younger workers. This tendency, shown clearly in Table 1, seems to indicate that older workers are introspectively more conscious of feelings of tiredness and readiness for work at specific hours of the work shift than are younger workers. It is possible, of course, that "objective" fatigue is equally present in the younger workers but that the younger workers are less affected in their subjective feelings by their continually changing organic conditions than are older workers. Another explanatory hypothesis is that younger workers simply have less insight into their feeling changes with successive hours of work.

Whatever the most tenable explanation, older workers in this study report significantly greater extremes of work feelings than do their younger associates.

Sex. Since all the manual workers are male, the sex comparison is limited to office personnel—48 males and 24 females—groups too small for any except suggestive comparisons. A suggestive tendency is present for female employees to report greater extremes of tiredness than do males.

Type of Work. Curves of work feeling throughout the work spell for manual work, office work, and supervising are displayed in Figures 1 and 2. It is significant that these three curves in each graph are all highly similar, despite the fact that they are derived from reports of employees doing dissimilar types of work and in different firms and industries. Introspectively, apparently, workers of widely differing types experience substantially the same feelings of tiredness and readiness for work at specific periods in the work spell. The similarity of these curves is all the more striking when it is considered that the manual workers are "swing" shift rather than regular day shift personnel. Extent of subjective tiredness feelings appears from Figure 2 to be in part a function of degree of manual effort involved in jobs performed. Supervisors

Table 1

Per Cent of Each Age Group Among Manual and Office Workers Reporting Feelings of "Most Tired" and "Most Rested" at Each Hour of the Work Shift and the Mean Deviations of Such Reports from Normal Expectancy

Hour of Work Shift	Per Cent "Most Tired"				Per Cent "Most Rested"				Chance Expec- tancy
	Manual		Office		Manual		Office		
	Age 20-35	Age 36-65	Age 20-35	Age 36-65	Age 20-35	Age 36-65	Age 20-35	Age 36-65	
1	32	20	27	41	27	38	31	6	25
2	5	7	3	18	36	29	51	53	25
3	16	18	18	0	20	15	15	35	25
4	47	55	52	41	17	18	3	6	25
5	100	100	100	100	100	100	100	100	100
6	23	16	33	12	28	33	25	24	25
7	15	2	9	12	35	41	42	52	25
8	25	37	29	35	24	14	24	12	25
	37	45	29	41	13	12	9	12	25
MD	100	100	100	100	100	100	100	100	100
N	10.3	15.5	11.2	14.5	6.5	10.3	12.3	16.3	0
	84	148	55	17	84	148	55	17	

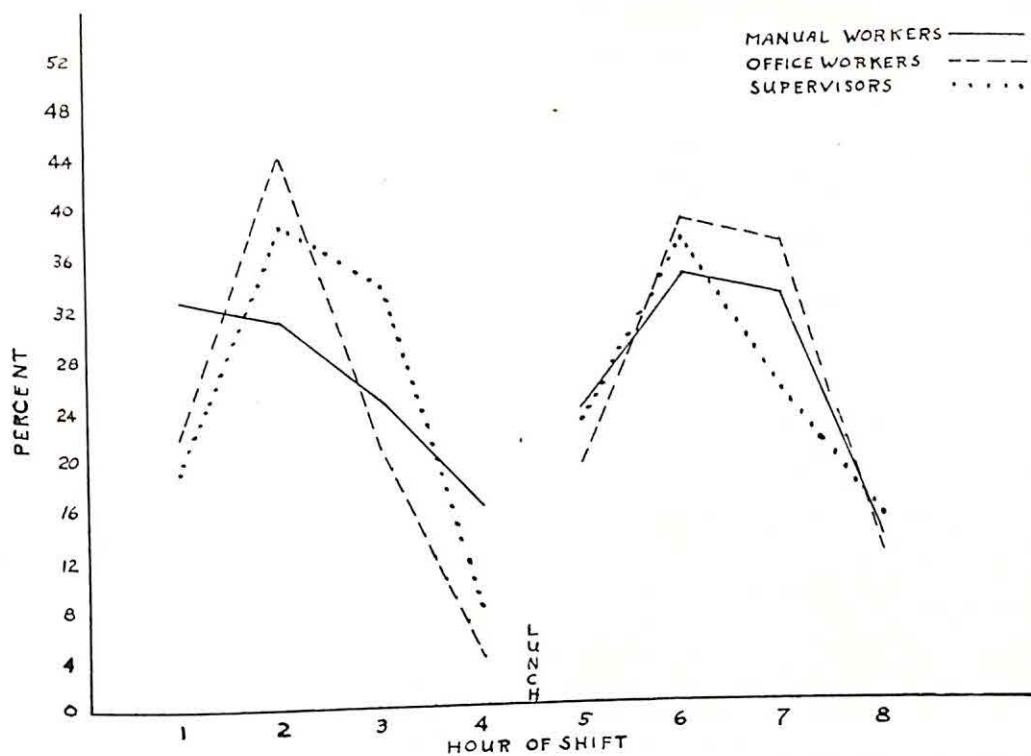


FIG. 1. Per cent of manual, office, and supervisory employees reporting maximal feelings of restfulness at each hour of each half of the eight-hour work shift.

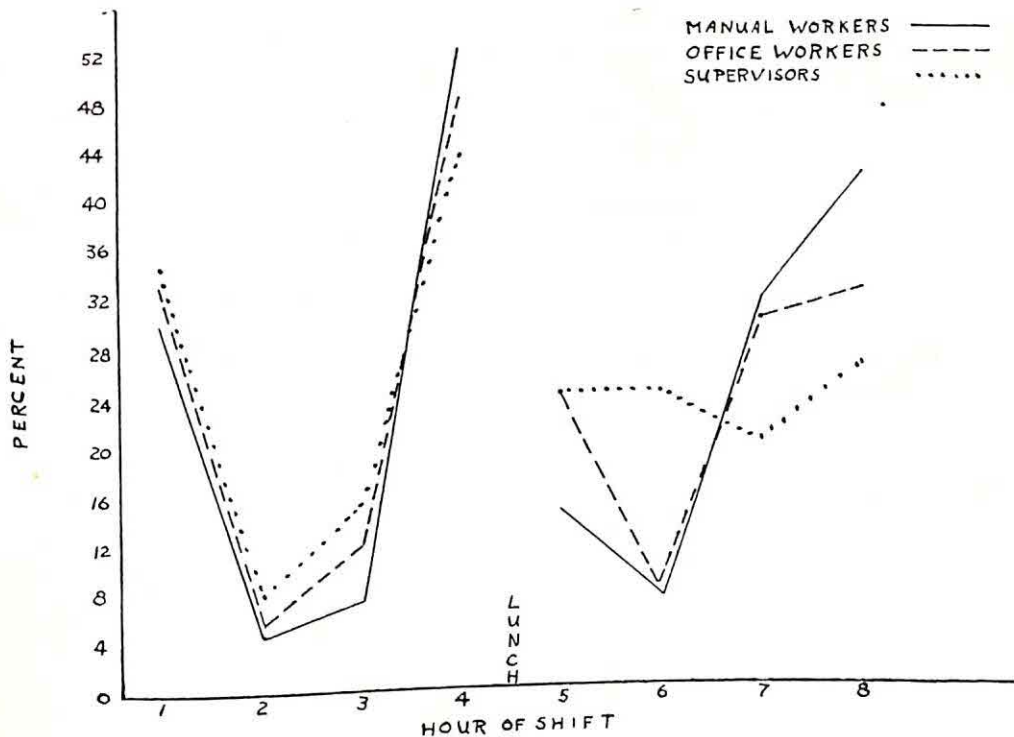


FIG. 2. Per cent of manual, office, and supervisory employees reporting maximal feelings of tiredness at each hour of each half of the eight-hour work shift.

show minimal variation about the line (25 per cent) of chance expectancy while manual workers show maximal variation for the three groups studied. Period of maximal tiredness seems to be the hour preceding the lunch period, while another peak of tiredness is during the last hour of the work shift. Readiness for work in terms of per cent of personnel reporting themselves as "most rested" is maximum in the second hour after the beginning of each work spell and it is minimal in the last hour of each work spell.

Summary

Manual, office, and supervisory employees totalling 379 from three different establishments were measured with a Kerr "tear ballot" for subjective feelings of tiredness and restfulness in the various hours of the eight-hour work shift.

1. Manual, office, and supervisory personnel report significantly differential feelings of tiredness or restfulness for various periods in the work shift.

2. Older workers report significantly greater variation of such feelings than do employees under age 36.

3. Curves of tiredness feeling and restfulness feeling throughout the work shift are remarkably similar for the manual, office, and supervisory employees in this study. The similarities are more impressive than the dissimilarities.

4. Maximal subjective fatigue is reported in the fourth and eighth hours of the eight-hour shift.

5. Maximal restfulness feeling is reported in the second and sixth hours of the shift, the second hour of each four-hour work spell.

6. In possible future evaluation of the psychological and efficiency advisability of the six-hour day, it is recommended that such data as these reported here be obtained on employees now engaged in six-hour shifts. Such new data should be examined particularly for: (a) less variability of tiredness feeling response; and (b) relative absence of high tiredness peaks just before the middle and end of the work shift.

Received September 16, 1949.

References

1. Goldmark, M. D., Hopkins, P. S. F., and Lee, F. S. Studies in industrial physiology: fatigue in relation to working capacity: comparison of an eight-hour plant and a ten-hour plant. *U. S. Public Health Service, Public Health Bulletin No. 106*, 1920.
2. Kerr, W. A. Where they like to work; work place preference of 228 electrical workers in terms of music. *J. appl. Psychol.*, 1943, **27**, 438-442.
3. Muscio, B. Is a fatigue test possible? *British J. of Psychol.*, 1921, **12**, 31-46.
4. Rothe, H. F. Output rates among butter wrappers: I. Work curves and their stability. *J. appl. Psychol.*, 1946, **30**, 199-211.
5. Ryan, T. A. *Work and effort*. New York: Ronald Press Co., 1947.

Accident Proneness of Factory Departments *

Willard A. Kerr

Illinois Institute of Technology

The extent to which individual accident proneness exists or has been a determinant of physical casualties in industry plainly has been exaggerated by many earlier authoritative writers according to more recent evidence (3, 7). Much factory data which appear at first examination to indicate that certain employees are persistent "repeaters" and therefore "accident prone" fail to substantiate such conclusion upon detailed probability study. Accidents distributed by chance (under the theory that a certain approximate number are inevitable under the existing total work situation in a factory department) will supply some workers with no accidents, some with one, some with two, and a few with even three or more (7). Because such analysis actually does succeed in most factory experience in explaining much of the individual employee "repeat" accidents data, the time-honored approach of the psychologist and psychiatrist (4) which emphasizes identification of subtle personality conditions which predispose to accidents by some employees seems to be a less promising approach than that which emphasizes study of the total psychological climate in which the typical employee of a group works. If proneness (or liability) to accidents exists such tendency may be a group psychological phenomenon as well as an individual psychological phenomenon.

The fact that intelligent safety engineers and industrial training personnel working with individuals and equipment often are unable to take some factory departments out of the "accident prone" column even after years of intense effort is proof that many group (as well as individual) psychological conditions may be operating.

The Present Study

Subjects for this study were 53 accident prone and non-accident prone departments in

the Camden Works of RCA involving 12,060 employees. These data were collected in 1943. Forty other variables were investigated in each department.

Accidents per hundred workers per year for these departments ranged from 0.0 to 22.7, although 38 of the departments had rates of less than four accidents per 100 workers. Severity ratings, based largely on days lost from work, also were obtained for each department with the advice of the plant safety director. These severity values ranged from 0 to 75.

It is only because of the grave importance of the objective that such unpromising potential correlates of accidents as some of these reported in this study were investigated. Of the forty variables studied only a few were significantly related to accidents, as expected, yet at least two of these results have not been reported previously in accident literature; therefore, they may justify the entire investigation.

Because both accident variable distributions were positively skewed and several of the variables studied consisted of dichotomous or two-interval data, the tetrachoric coefficient of correlation (2) was employed. The statistically significant correlations (five per cent level) in Table 1 are indicated according to use of Kelley's reliability formula (6) and the Guilford-Lyons tables (5).

Inspection of these significant correlations reveals that accidents tend to occur with greatest frequency in those factory departments with lowest intra-company transfer mobility rates, smallest per cent of employees who are female and on salary, least promotion probability for typical employee, and highest mean noise level.

While departments highest in accident frequency usually also are above average in accident severity, the severe accident departments have some systematic characteristics which are found less often in the high accident departments. High severity departments are heavily

* Acknowledgment of invaluable advice and assistance in obtaining the accident data is made to O. C. Boileau, Safety Department, Radio Corporation of America, and to Dean F. H. Kirkpatrick, Bethany College, for constructive criticism.

Table 1

Correlations between the Frequency and Severity of Accidents in 53 Factory Departments and Each of 40 Other Variables in a Large New Jersey Factory *

Variable	Accident Frequency	Accident Severity
1. Number of production employees	.36	.16
2. Total employees	.18	-.12
3. Per cent of employees who are male, production	.24	.63
4. Per cent of employees who are male, salary	.20	.50
5. Per cent of employees who are production workers	.28	.15
6. Production employees per supervisor	.16	.08
7. Mean hours worked per week per production male	.12	-.29
8. Mean hours worked per week per production female	.20	-.20
9. Mean base pay of production males	-.07	.22
10. Mean base pay of production females	.07	-.18
11. Sex hours differential, mean	.24	-.28
12. Sex wage differential, mean	-.31	-.14
13. Intra-company transfer mobility	-.44	.28
14. Sex-ratio imbalance	.21	.51
15. Gross turnover rate (including accessions)	-.12	-.30
16. Avoidable turnover rate (including accessions)	-.06	.02
17. Avoidable separation rate	-.12	-.30
18. Per cent of employees who are salaried male	-.16	.20
19. Per cent of employees who are salaried female	-.40	.13
20. Per cent membership in company athletic association	-.08	.05
21. Accident frequency	xx	.64
22. Accident severity	.64	xx
23. Efficiency (plant manager rating, three-month period)	.06	-.21
24. Efficiency (mean rating of ten competent judges)	-.09	.16
25. Mean job security (mean rating of twelve competent judges)	-.05	.02
26. Mean supervisory quality (mean rating, twelve judges)	-.22	.08
27. Mean job prestige (mean rating, twelve judges)	-.30	.03
28. Mean promotion probability (mean rating, twelve judges)	-.40	-.50
29. Mean job monotony (mean rating, twelve judges)	.13	.03
30. Degree of completion of work (rating, suggestions supervisor)	-.07	.05
31. Fertility of suggestion field (rating, suggestions supervisor)	-.11	-.17
32. Suggestion quota (established by suggestions supervisor)	.20	.12
33. Total suggestions submitted	.00	-.40
34. Per cent of suggestion quota met	.00	-.54
35. Per cent of suggestions adopted	-.37	.16
36. Wage incentive system	.12	-.35
37. Mean noise level	.00	.13
38. Labor-management mean morale rating (mean of 39 and 40)	.42	-.40
39. Morale as rated by personnel manager	.00	-.36
40. Morale as rated by union local officers (pres. and vice-pres.)	.17	-.35
41. Youthfulness of employees (per cent under 26)	-.23	-.57
42. Tenure (per cent employed more than twelve months)	-.37	.55
	.36	

* Coefficients in bold face are statistically significant at the five per cent level or better.

male in sex ratio for salary as well as production personnel; they are low in mean promotion probability, low in fertility of suggestion field, low in employee suggestions contributed, high (relatively) in average employee age level, and higher in average employee tenure.

Most of these correlations undoubtedly are artifactual rather than causal in significance. However, a few are worthy of further study and interpretation. Possibly substantial intra-company transfer mobility makes employees more alert and interested in their work en-

vironment, resulting in fewer accidents. The cross-fertilization of ideas which probably accompanies intra-plant mobility may act also to reduce accident hazards and promote positive cooperation with safety personnel.

The tendency for departments lowest in promotion probability to be high in both accident severity and frequency may be of considerable psychological significance. It is plausible that when promotion is too unlikely, the *typical* employee may develop accident prone *attitudes of relative indifference to the work environment*. A reasonable chance to get ahead may constitute an incentive which not only stimulates the employee to do better work but may make him more alert to avoid hazards which may detain him in his progress.

Accident prone departments usually have above average noise levels. Whether the noise level is causal of accidents or merely an incidental correlate of hazardous factory operations is not entirely clear; it appears to be both causal and incidental. Certainly the reduction of excessive noise levels whenever such reduction is practicable can be expected to do more good than harm as regards accident records.

The pattern of correlates of accident severity is somewhat different from that of accident frequency. As might be expected, maleness is a marked characteristic of severe-accident departments; probably females are rarely placed on the most dangerous jobs or in the most "strenuous" departments. Also, male employees tend to be older; Chaney and Hanna (1) found that the probability of fatal or disabling results is greater among older than among younger accident cases.

Less easily explained, however, is the fact that severe accident departments are units which tend to show a poor performance in contributing to the plant suggestion system. Superficially, it appears that departments which lag in making constructive suggestions through the employee suggestion boxes lag also in correcting dangerous conditions and in passing tips around on how not to get hurt; the superficial interpretation easily may be the valid one.

Another tenable hypothesis is that the average "foresight factor" of intelligence is lower in the severe accident departments because

foresightful employees tend to avoid or transfer away from dangerous work departments. A third hypothesis, also possibly tenable, is that severe accident departments are those which have been so highly systematized and perfected from the industrial engineering standpoint that the average worker feels no incentive to try to improve the work or workplace through either employee suggestion boxes or alertness to unexpected hazards. While this latter hypothesis is improbable, it does seem highly significant that departments which are high in suggestion fertility are low in accident severity.

A New Frame of Reference for Safety Promotion

Perhaps, as some of the correlations in this study seem to suggest, a fundamental change in the total psychological frame of reference in which the *average* employee works is the basic key to reduction of industrial accidents. *This probably can produce the probability that fewer total accidents will happen.*

A psychological work environment that rewards the worker emotionally for being alert, for seeking to contribute constructive suggestions, for passing a tip to a co-worker on how best to do something or how not to get hurt appears from this research to be a profitable goal to work toward. Creating or promoting such an environment undoubtedly calls for a much broader perspective of approach to accidents than has hitherto been considered by most managements.

An additional item of circumstantial proof that that which promotes alertness also tends usually to minimize accidents is the fact in this study that the departments with incentive-pay systems have no more accidents than other departments. Approximately half of the departments studied are on incentive systems. These same departments are "problem" departments in many respects (higher turnover, more monotonous work, less job prestige, and less promotion probability). In fact, they have almost all undesirable characteristics except accidents in greater quantity than do the non-incentive departments. The "normal expectancy" record as regards accidents in incentive departments practically is in defiance of

physical and even some psychological work conditions. Even though incentive systems rarely succeed as much as they theoretically should in motivating the worker, they nevertheless appear to make him more alert to attain a reasonable productive goal and this alertness apparently makes him safer in his operations. These observations on incentive systems are, of course, somewhat speculative. However, the need for providing emotional rewards for alertness seems highly probable from this research. Such rewards might include economic rewards, prestige-building honors, extra privileges, and representation on special committees and councils. These rewards held as attainable goals by workers in "dead end" jobs should operate to raise the average level of alertness, not just to hazards but to everything.

Summary

Accident severity and frequency were correlated with each of forty other variables in the 53 departments of an electronics factory.

1. Accident frequency is associated with low intracompany transfer mobility, small per cent of employees who are female and on salary, low promotion probability, and high noise level.

2. Accident severity is associated with predominant maleness, low promotion probability, low fertility of suggestion field, low suggestions record, non-youthfulness of employees, and high average tenure of workers.

3. A common explanatory factor among the accident frequency correlates appears to be *depressants to alertness*. The same factor appears to be present in most of the severity correlates.

4. Industry should direct increased attention toward enlivening of the psychological work environment, particularly with reference to provision of more and more emotional reward goals as incentives to raise the average level of alertness.

Received August 15, 1949.

References

1. Chaney, L. W., and Hanna, H. S. The safety movement in the iron and steel industry. *Bur. Labor Statistics Report* 234, 1918.
2. Chesire, L., Saffir, M., and Thurstone, L. L. *Computing diagrams for the tetrachoric coefficient of correlation*. Chicago: Univ. of Chicago Bookstore, 1933.
3. Cobb, P. W. The limit of usefulness of accident rate as a measure of accident proneness. *J. appl. Psychol.*, 1940, **25**, 154-159.
4. Dunbar, Flanders. Medical aspects of accidents and mistakes in the industrial army and in the armed forces. *War Medicine*, 1943, **4**, 161-175.
5. Guilford, J. P., and Lyons, T. C. On determining the reliability and significance of a tetrachoric coefficient of correlation. *Psychometrika*, 1942, **7**, 243-249.
6. Kelley, T. L. *Statistical method*. New York: Macmillan, 1924.
7. Mintz, A., and Blum, M. J. A re-examination of the accident proneness concept. *J. appl. Psychol.*, 1949, **33**, 195-210.

The Rank-Comparison Rating Method

Reign H. Bittner

The Prudential Insurance Company of America

and

Edward A. Rundquist

Personnel Research Section, AGO

Validating tests or other measures of aptitude for an industrial job always begins with a search for an adequate criterion. A variety of measures have been tried as the yardstick of job success: turnover, absenteeism, medical visits, production records, merit ratings and specially devised ratings. Turnover, absenteeism and medical records may be adequate criteria for validating personality tests or application blank material. These types of criteria will not be discussed here as our concern is with predicting ability to learn to do a job.

Production measures would appear to be the ideal criterion for validating tests of aptitude. Unfortunately, however, production is affected by so many things not under the control of the individual that such records rarely reflect accurately what an individual can do. They are affected by variation in the quality of the material handled, by the pace of the machine, by the pace of others in the work group, by the correlation of job assignment with length of service, and even by the correlation of job assignment with ability to do the job. These and other factors operate almost universally to make production records inadequate as a criterion of ability to perform a job. Even when such records are collected over a long period of time the factors mentioned do not necessarily average out.

Merit ratings are often available as a possible yardstick of job success. However, upon examination they usually turn out to be of little use. *First*, merit rating procedures usually have several purposes only one of which is to get a measure of the person's ability to do the job. Other purposes such as aiding the supervisor to deal more effectively with his people, building morale, etc., affect the ratings in ways making them undesirable as criteria. *Second*,

the distribution of merit ratings is often so skewed that their use as a discriminative criterion is ruled out. *Third*, most merit ratings are collected under such uncontrolled conditions and with so little training of the raters that they become valueless as criterion measures. It should be noted, however, that these defects are not necessary as both Ferguson (5) and Mahler (7) have shown.

Specially devised ratings are commonly used as criteria since merit ratings and production measures cannot be relied upon to furnish a good measure of individual performance. These special ratings have been of many kinds. Probably most common are the variations of the rank order or paired-comparison methods. The writers have developed the rank-comparison rating method, a new method of obtaining a rating criterion which has proved both practical and valuable in many industrial researches. It is the purpose of this paper to describe the rank-comparison rating method and to present data concerning certain characteristics of ratings obtained by this method.

The Rank-Comparison Rating Method

The rank-comparison rating method combines features of the ranking and paired-comparison methods. It involves the following general steps:

1. Separation of the total group into random sub-groups.
2. Ranking within sub-groups.
3. Successive merging of sub-groups by a modified paired-comparison method.

The end result is a ranking of the total group from best to poorest achieved without the laborious comparisons involved when large groups are handled by the straight forward paired-comparison method or the confusion that arises in trying to rank a large group.

Preliminary Preparation

The preliminary preparation required before contacting the rater is as follows:

1. Prepare a small name card for each person to be rated.
2. Arrange the cards in alphabetical order.
3. Divide the rater's total cards at random into two or more sub-groups with from 15 to 20 cards in each sub-group, keeping the groups as nearly equal as possible. Decide how many sub-groups are needed. Then take the alphabetically arranged pack of cards and deal out the required number of groups as in dealing hands in a card game, dealing one card at a time.

It is convenient, though not necessary, to divide the total group into 2, 4, or 8 sub-groups. This makes it possible to work with the same number of cards in each group at each stage of the merging process.

Obtaining the Ratings

The ratings are obtained during a conference with the rater. The procedure is as follows:

Initial Judging

1. Explain carefully and exactly to the rater what is to be considered in judging the persons to be rated.
2. Lay out the cards for the first sub-group alphabetically in front of the rater in one or two columns depending on the size of the group.
3. Ask the rater to choose the *best* person in the group, emphasizing again the basis on which the choice is to be made.
4. Place the card of the person chosen as *best* at the top of a new column of cards.
5. Ask the rater to choose the *poorest* person in the group.
6. Place the card of the person chosen as *poorest* at the bottom of the new column of cards.
7. Ask the rater to choose the *best* person remaining in the group. Place this person's card in the new column of cards *under* the one previously placed at the top of the column.
8. Ask the rater to choose the *poorest* person remaining in the group. Place this person's card in the new column of cards *above* the card already at the bottom of the column.

9. Continue this process, alternately choosing *best* and *poorest* persons among the cards remaining in the original group and building the new column by placing each card selected as *best* *under* the last card placed at the top of the column, and each card selected as *poorest* *above* the last card placed at the bottom of the column. It should be obvious at this point that the cards are being placed in rank order by building from both ends toward the middle.

10. When all cards have been transferred to the new column by this alternating selection procedure, ask the rater to check over the way the persons have been ranked making any adjustments considered necessary. Permit persons to be moved up or down in rank if the rater so desires. Ordinarily, few adjustments will be made.

11. Pick up the cards, keeping them in rank order from best of all on top to poorest of all on the bottom. Put this pack of cards aside for the present.

12. Repeat steps 2-11 for each of the other sub-groups.

Merging

13. The next steps in the procedure involve the merging of the sub-groups so that the end result is the total group ranked in order from *best-of-all* to *poorest-of-all*. The merging procedure will be described for any two sub-groups. If there are more than two sub-groups, the steps in merging are shown in Table 1 for any number of groups up to six. For example, if there are three sub-groups, merge sub-groups 1 and 2, and then use the same procedure to merge sub-group 3 with the total of sub-groups 1 and 2.

14. Place the cards for sub-groups 1 and 2 before the rater in two stacks, each stack arranged in rank order from *best* person on top to *poorest* person on the bottom. The rater is directed to consider the *two persons whose names show on top of the two stacks of cards*. Ask the rater, "Which of these two is the better?" The card chosen is removed and placed *upside down* on the table to begin a new stack of cards.

15. Ask the rater, "Which of the two now showing is the better?" The card chosen is removed and placed *upside down* on the new stack.

Table 1
Showing Steps in Merging

Steps in Merging	Groups to Merge at Each Step with Varying Numbers of Sub-Groups *				
	2 Sub-Groups	3 Sub-Groups	4 Sub-Groups	5 Sub-Groups	6 Sub-Groups
A	G1 with G2	G1 with G2	G1 with G2	G1 with G2	G1 with G2
B		G(1+2) with G3	G3 with G4	G3 with G4	G3 with G4
C			G(1+2) with G(3+4)	G(3+4) with G5	G5 with G6
D				G(3+4+5) with G(1+2)	G(1+2) with G(3+4)
E					G(1+2+3+4) with G(5+6)

* Sub-groups identified as G1, G2, G3, etc.; G(1+2) is the combined group after sub-groups 1 and 2 have been merged, etc.

16. Continue the procedure, each time asking the rater to choose the better of the two names showing and placing the card chosen upside down on the new stack of cards. Usually, the question need not be asked more than twice; the rater quickly grasps the idea and proceeds through the comparisons without need of further prompting. When all cards have been transferred to the new stack, the two groups will have been merged into a single combined group which is ranked in order from *best* to *poorest*.

17. Continue merging groups according to the merging steps shown in Table 1. In merging each pair of groups, repeat the procedure given in 14, 15, and 16 above.

Final Check

18. At the conclusion of the merging process, the total group will have been placed in rank order from *best* to *poorest*. As a final check on the rank order, lay out the cards in rank order before the rater and ask him to make any adjustments felt to be necessary. The rater is permitted to move persons up or down in rank if desired. Experience has shown that few adjustments are made, but it is desirable to give the rater this opportunity.

Statistical Treatment of the Ratings

The procedure results in rank-order ratings which are more amenable to statistical treatment if converted to another scale. In our

studies, it has been found most useful to convert the rank-orders to standard scores. The method described by Garrett (6) has been used in making this conversion.

Characteristics of Rank-Comparison Ratings

Reliability. The rank-comparison ratings are quite reliable in the sense that the same raters given the same directions will give essentially the same ratings even after a considerable lapse of time. For example, a foreman rated 75 factory women and rerated them three months later. The correlation between the two ratings was .92 (11). Another foreman rated and rerated 31 factory women one month apart. The correlation between the two ratings was .89 (2).

The importance of measuring reliability of ratings under the same conditions is illustrated in the first study cited above. In an effort to get raters to control their bias favoring long-service employees, a straight paired-comparison method was tried along with special efforts to make the rater discriminate between ability and length of service of the people rated. The correlation between ratings by the rank-comparison method and the paired-comparison method was .26. This is a marked change from the .92 when the same method with the same directions were used. Other foremen changed as noted in the study, but not to the same degree as this particular one.

Agreement Among Raters. Agreement among raters is more likely to be a function of factors other than the method of rating. The degree of knowledge each rater has of the people rated and the varying standards used by the raters in judging performance are particularly important factors. It is of some interest, however, to note the degree of agreement achieved when the rank-comparison method is used. Forty-eight factory women were rated by their foreman, assistant foremen and by nine inspectors (9). Each inspector rated from 4 to 40 of the women. The inter-correlations between ratings were: foreman and assistant foreman +.65; foreman and "average inspector" +.67; and assistant foreman and "average inspector" +.73.

In another study, 97 supervisors and foremen were rated by three raters on personality

suitable for supervision (4). The intercorrelations between raters were: plant manager and personnel director +.63; plant manager and training director +.45; and personnel director and training director +.53.

Agreement among the raters in these two studies is moderate to fairly high. Comparable results have been found in other studies. This agreement can in no sense be interpreted as reliability of the ratings. However, if there were no agreement among raters who really knew their people and who were reasonably consistent in their rating standards, it would raise questions concerning the adequacy of the method.

Relation to Paired-Comparison Ratings. A completely controlled experimental comparison of the rank-comparison method with the paired-comparison method is not available. Two studies are available from which inferences as to the relationship between the two methods can be drawn. The results of the two studies indicate that if all conditions are controlled the two methods give essentially the same results.

A study previously cited (11) involved rating of four groups of factory women by the rank-comparison method and rerating seven months later with a partial paired-comparison technique. In the rerating technique, the sub-groups set up in accordance with the rank-comparison method were ranked by a straight paired-comparison technique after special directions were given emphasizing that the rater should discount as much as appropriate the length of time persons had been on the job and rate solely on actual ability to perform the job. (These special directions were not given in the original rank-comparison rating.) The sub-groups were then merged according to the rank-comparison technique. The rating and rerating correlations for four raters are: Rater 1 (N=75) +.26; Rater 2 (N=82) +.90; Rater 3 (N=64) +.84; and Rater 4 (N=67) +.70. The correlations between the two ratings are high in all but one case. The low correlation for Rater 1 appeared on investigation to be due to the change in directions on evaluating length of time on the job rather than to the paired-comparison technique used in ranking the sub-groups. It would seem then that to the extent

the paired-comparison technique was involved it did not markedly change the ratings.

A second study (1) involved rating of 18 factory women. The rank-comparison method was modified slightly as follows: sub-groups were ranked by their foremen and the department supervisor then merged the sub-groups in the standard way. The department supervisor several days later then rated the 18 women by a straight paired-comparison technique. The correlation between the rank-comparison and the paired-comparison ratings was .97. Obviously, there is no essential difference in the ratings obtained by the two methods.

Relation to Rating Scale Ratings. A study in the selection of supervisors (4) involved the use of several types of criteria of personality suitable for supervision: rank-comparison, a 9-trait rating scale with each trait rated on a 5-point scale, and a 2-point (above and below average) rating scale on overall personality for supervision. Two raters rated by rank-comparison and the 2-point overall rating methods all of the 96 supervisors they knew well enough to rate. In addition, they rated the supervisors immediately under them on the 9-trait scale. The correlations between the rank-comparison and the other two types of ratings for each rater are: Rater 1, rank-comparison vs. two-point overall scale ($N=92$), bi-serial $r=.93$; Rater 1, rank-comparison vs. nine trait scale ($N=19$) $=+.88$; Rater 2, rank-comparison vs. two-point overall scale ($N=96$), bi-serial $r=+.75$; and Rater 2, rank-comparison vs. nine trait scale ($N=13$) $=+.77$.

The rank-comparison method gives results closely comparable to the other two methods for Rater 1 and quite comparable for Rater 2. Depending on the type of rating desired from other considerations, it is indicated that the rank-comparison method may be used with confidence instead of rating scale methods.

Relation to Production Criteria. Rank-comparison ability ratings and production criteria should be closely related if both are valid and reliable measures. The usual inadequacies of production criteria mentioned earlier were all present in such criteria available in our studies. Thus, a close relation between the two types of criteria was not expected. However, some evidence has been found to indicate that as the

worker's individual control over production increases, the relationship increases between production measures and the ratings. In one study where the individual's production was largely machine-paced and where there was no opportunity for differences in type of material handled to cancel out, the correlation between the production criterion and the foreman's rating was .24 (10). In another situation the production of the individual was considerably more under the worker's control (9). In this situation where each worker could set her own pace, correlations between ratings and average pay period efficiency was .84 for the inspector's rating, .73 for the assistant foreman's rating, and .78 for the foreman's rating. In a third study where again production was fairly well controlled by the worker, the correlations between supervisors ratings and three production measures were as follows: .66 with production speed; .50 with production quality, and .70 with overall production efficiency (1). These findings suggest that the rater makes allowances for difficulty of job assignment, quality of products handled and the like in making his ratings. It also suggests that with proper care, ratings can be obtained which will be of almost as great value as adequately controlled production measures for the purpose of validating tests. This is an important consideration where adequate production records are not routinely kept since ratings can be obtained much more cheaply.

Predictability of Rank-Comparison Ratings. It might be expected that ratings would be more predictable than production records if raters make allowances for the factors that often operate to make production records inadequate criteria of the individual's worth on the job. There is some evidence that this is the case. In the first study cited above (10) involving 63 factory women, a test battery was developed that correlated .26 with the production measure. However, it was possible to develop a test battery that correlated .47 with the rating criterion. In the second study cited above (9), 37 women were tested. These were divided into two groups, one of 20 women with less than ten months service, and one of 17 women with ten months or more of service. The test battery was developed on the basis of an average production and rating criterion.

Table 2
Correlations between Test Battery and Various Criteria

Criterion	Test-Criterion Correlations		
	Short Service Group (N = 20)	Long Service Group (N = 17)	Total Group (N = 37)
Ave. Production and Rating	.61	.44	.49
Ave. Foreman and Ass't Foreman Rating	.70	.45	.49
Production Efficiency*	.50	.38	.41
Ave. Inspector's Rating	.47	.42	.29

* The average of 10 to 26 pay-period efficiency indices based on standards set by Industrial Engineering Department.

The correlations of the test battery with the various criteria for the short and long service groups are shown in Table 2.

The most predictable single criterion is the average rating of the assistant foreman and the foreman. Efficiencies, however, tend to be more predictable than the inspectors' ratings. Perhaps the inspectors whose job is to detect violations of quality standards are overly influenced by the knowledge of defects found and do not make sufficient allowances for the difficulty of the job or for varying lengths of service. They seem confused by the length of service factor. Instead of a correlation for the total group somewhere in between those for the short and long service groups as was found for the other criterion measures, the correlation for the inspectors is lower for the total group.

Advantages of the Rank-Comparison Method

Traditionally, ranking methods have been considered as a substitute for the paired-comparison method. From the evidence presented above, the rank-comparison combination of the two methods appears to yield as satisfactory a result as the paired-comparison method. Moreover, it has four great operational advantages: (1) it is easily understood by the raters; (2) raters like the method and have confidence in it; (3) it can be applied to large groups; and (4) it requires very little of the rater's time.

The method is easily understood by the rater. At least 50 department heads, foremen and inspectors have rated their people with this method. Many of these raters had little ver-

bal facility but no difficulty in comprehension has been encountered. In fact, the method of merging the ranked sub-groups was invented when one rater was unable to understand a system whereby more than two names were exposed simultaneously. Once the present method had been devised, however, the rater proceeded with no further difficulty.

Raters like and have confidence in the method. Their reaction has always been favorable. Not only is the method liked by the raters but it invokes in them a feeling of confidence that the ratings are accurate measures of ability. Many have commented that this method should be used in giving merit ratings, although the writers do not agree with this.

The method can be used with large groups. When groups of 100 or more are to be rated, the paired-comparison method obviously becomes too unwieldy. With 100 cases it would involve 4,950 comparisons. It is also difficult to attempt to rank this many people. To overcome this difficulty, large groups are often divided into random sub-groups of equal size and then either the rank order or paired-comparison method is used within sub-groups. This involves the assumption that sub-groups are truly random groups. The rank-comparison method gets away from such problems and such assumptions. It does not matter how large the group is; it can be ranked from 1 to N by this method with very little complication. The feature that makes this possible is that the rater never has to consider more than 20 persons at one time. In practice, it was discovered that when more than 20 cards

were placed before the raters, they sometimes became confused and were hesitant in making the original rankings. With 20 or fewer cases, no difficulty has been encountered.

The method requires very little of the rater's time. While no systematic records have been kept, instances are known where close to a 100 people have been ranked in less than 15 minutes.

Conclusion

Ratings, with all their difficulties, are the most common criterion data in industrial personnel research. Not only are they the most common but they are probably the best under most circumstances. Production records, if obtained under conditions of equal training and control of all factors affecting individual performance except individual differences in ability, would be the ideal criterion. With the rare exception of some training programs (8), such conditions are seldom met.

The rank-comparison method of rating presented here is a practical and useful method which has certain advantages over other methods. It does not, however, solve all the problems incident to the use of ratings as validation criteria. Since in the majority of industrial situations the experimenter is reduced to the use of ratings, it would seem useful

to devote more study to methods of obtaining these ratings. The solution of these problems awaits further research.

Received August 12, 1949.

References

1. Bittner, R. H. The selection of bottle decorating machine operators. (Unpublished study.)
2. Bittner, R. H. The selection of bottle inspector-packers. (Unpublished study.)
3. Bittner, R. H. The selection of handyman-inspectors. (Unpublished study.)
4. Bittner, R. H., and Rundquist, E. A. Development of a supervisor personality test. (Unpublished study.)
5. Ferguson, L. W. The effect upon appraisal scores of individual differences in the ability of superiors to appraise subordinates. *Personnel Psychol.*, 1949, 2, 377-382.
6. Garrett, H. E. *Statistics in psychology and education*. New York: Longmans, Green and Co., 1937. Pp. 168.
7. Mahler, W. R. An experimental study of two methods of rating employees. *Personnel*, 1948, 25, 211-220.
8. McGehee, W. Cutting training waste. *Personnel Psychol.*, 1948, 1, 331-340.
9. Rundquist, E. A. Predicting success of glass selectors. (Unpublished study.)
10. Rundquist, E. A. The selection of tumbler inspector-packers. (Unpublished study.)
11. Rundquist, E. A., and Bittner, R. H. Using ratings to validate personnel instruments: a study in method. *Personnel Psychol.*, 1948, 1, 163-183.

Validity of an Objectivity Key on a Short Industrial Personality Questionnaire *

Edward R. Carr and Harold F. Rothe

Stevenson, Jordan and Harrison, Inc., Chicago, Illinois

In a previous paper, Rothe described the use of an Objectivity key on a short industrial personality questionnaire (3). The Objectivity key described there and referred to in this paper consists of six items patterned after the L scale of the MMPI (1). The Objectivity key was shown to permit the adjusting of scores on some other keys so that highly objective persons would be compared with norms based upon other highly objective persons; persons of medium or low objectivity would also be compared with their appropriate groups. This technique is intended to minimize the effects of "faking" on questionnaires.

This technique is only valuable, however, when the Objectivity key is valid, and when the other keys are valid. A valid Objectivity key is one that indicates which persons are being objective and which ones are not being objective while answering the questionnaire. A non-objective person is one who is "putting his best foot forward" and attempting to "look good" on the questionnaire. A highly objective person is one who is extremely frank while answering the questionnaire, not attempting to hide what are apparent faults or weaknesses.

The purpose of the present paper is to present some data that indicate that the Objectivity key used in this study does separate highly objective from non-objective respondents.

Experimental Technique

The industrial personality questionnaire (3) was administered three times to a group of fifty college students. Different instructions were given orally to the group each time. The first instructions were to "fake the questionnaire" so as to "look good" for a job for which they were to assume they were applying. The second instructions, given after all stu-

dents had finished the form (in about five minutes), were to "fake the form to look bad." The third instructions, given after all students had finished the form the second time, were to "be as honest as possible."

As far as could be observed, good rapport existed between administrator and students, and it was believed that the students were cooperating. The students were freshmen, chiefly in an engineering curriculum.¹ The technique of several administrations with different instructions is generally similar to the technique used by Giese and Christy, reported in Tiffin (4) and by Longstaff (2). The results (means and s.d.'s) for the four personality keys for the three kinds of test administration are shown in Table 1. The actual distributions of these scores are filed in ADI.²

Results on The Objectivity Key

As Table 1 shows, the responses of the students on the Objectivity scale varied from trial to trial and in the anticipated direction. That is, when told to "fake to look good" the students obtained very low Objectivity scores. When they faked to make themselves "look bad" they obtained very high Objectivity scores. When they were "honest" they obtained the pattern of Objectivity scores, resembling a normal curve, that is customarily obtained in samples of industrial personnel.

The critical ratio between "good" and "bad" was 17.27; between "good" and "honest" was 6.05; and between "bad" and "honest" was

¹ The writers wish to thank Dr. Ernest McCormick, Department of Psychology, Purdue University, for permitting this experiment to be conducted in one of his classes.

² Tables 2, 3, 4, and 5 have been deposited with the American Documentation Institute and may be obtained by ordering Document No. 2723 from American Documentation Institute, 1719 N Street, N.W., Washington 6, D. C., remitting \$.50 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$.50 for photocopies (6 × 8 inches) readable without optical aid.

* The authors wish to acknowledge the assistance of Miss Judy Yackle in preparing this paper.

Table 1

Means and S.D.'s of Four Personality Keys for Three Kinds of Test Administration

Note N = 50 engineering college freshmen.

Key	Mean	S.D.
Objectivity:		
Look good	1.1	1.5
Look bad	5.3	.9
Be honest	3.0	1.6
Emotional Score:		
Look good	3.6	2.1
Look bad	10.6	1.5
Be honest	5.2	2.4
Social Dominance:		
Look good	7.9	.9
Look bad	1.2	1.4
Be honest	6.1	2.2
Drive Scores:		
Look good	6.1	1.3
Look bad	4.7	1.2
Be honest	6.3	2.0

8.68. All of these differences are significant. It is apparent, then, that college students can vary their Objectivity scores, depending upon the instructions given them, and presumably depending upon their set. Since the changes in score are in the direction that would be expected, according to the theory underlying the Objectivity key, or the MMPI L-Scale, it may be concluded that the Objectivity key does provide a measure of the extent to which the respondents are "faking." A scale with a greater range would, of course, provide a more adequate measure.

Results on the Emotional Key

In the previous paper it was shown that respondents with low Objectivity scores tended to have low scores on the so-called Emotional key. Respondents with high Objectivity scores had high Emotional scores. Accordingly the Objectivity score was found to be useful in interpreting the Emotional score (i.e., by making adjustments upwards or downwards in the Emotional score).

The students' Emotional scores showed the same relationships (see Table 1). The students, when "looking good," were being non-objective in order to "look good," and showed low Emotional scores. The mean score was

3.6. When the students were "looking bad," the mean Emotional score was 10.6. When they were "honest," the mean was 5.2. The differences between these three conditions are again all significant, the critical ratios being 19.55 between "good" and "bad," 3.51 between "good" and "honest," and 13.75 between "bad" and "honest." It is, therefore, apparent that college students can vary their Emotionality scores, depending upon their instructions and presumably upon their sets. It may also be concluded that low Emotional scores may be associated with low Objectivity scores, and high Emotional scores may be associated with high Objectivity scores.

There is a possibility that there is an inevitable relationship between Objectivity and Emotional scores and that both are measuring the same thing. Data to be published shortly rule out this possibility. Leaving that question for the moment, it has been shown here that the Objectivity key measures what it purports to measure and that the technique of interpreting Emotional scores within a frame of reference established by the Objectivity scores is a valid one.

Results on the Social Dominance Key

Social Dominance scores were not found to vary with Objectivity scores for the industrial sample previously reported. Interestingly enough, for the college sample discussed in this paper, Social Dominance scores were found to vary with the administrative instructions, and also with the Objectivity scores.

The mean Social Dominance score, under the "look good" conditions, was 7.9; the mean when "looking bad" was 1.2; the mean when "being honest" was 6.1 (see Table 1). The critical ratios are: 27.86 between "good" and "bad"; 5.15 between "good" and "honest"; and 13.15 between "bad" and "honest"; and all of the differences are significant.

These results are particularly interesting in view of the fact that the students, when "honest" give essentially the same distribution of Social Dominance scores as does the industrial sample previously reported. That is, college students apparently conceive of "Social Dominance" as being a desirable set of habits

for jobs to which they might aspire ("look good"), but industrial personnel of various categories do not "fake" questionnaires to make themselves appear more highly socially dominant than they are. That is a highly interesting point deserving more research.³

Results on the Drive Key

The industrial sample previously reported did not show a variation in Drive scores with Objectivity scores. The college students reported here do show some relationship between these two keys.

The mean score on Drive when "faking good" was 6.1; the mean when "faking bad" was 4.7; and the mean when "honest" was 6.3 (see Table 1). The critical ratios are: 5.68 between "good" and "bad"; 0.42 between "good" and "honest"; and 4.73 between "bad" and "honest." The difference between "good" and "honest" is not significant but the other two differences are significant. Thus, the college students could "fake" the Drive key to show that they had little drive, but they could not, or did not, "fake" to show more drive than they actually possess.

There are two additional features about the college students' not "faking" a higher drive than they did in order to look good. One is that their Drive scores, while "honest," were higher than while faking to look good, although insignificantly so. This leads to the possibility that this key is too subtle to fake. Unfortunately, the writers have no data to answer that problem.

The second feature is the possibility that the students actually had such a high drive they did not believe they would have to fake it. However, the students did not have an unusually high distribution of Drive scores, being substantially the same as those of the industrial samples previously described.

Accordingly it is tentatively concluded that the Drive key is too subtle to be faked, with the exception of one or maybe two items that can be faked to show lack of drive.

³ Neither of the groups of persons knew the names of the keys, but the Social Dominance items are fairly obvious; i.e., "I dislike walking across the middle of a room full of people."

Summary

In a previous paper the use of an Objectivity key to establish a frame of reference for interpreting certain other keys on a short industrial personality questionnaire was described. The present paper describes an experiment to establish the validity of that Objectivity key. This experiment was conducted with college freshmen rather than with an industrial population. The experiment should be repeated with an industrial sample.

When instructed to "fake" the questionnaire to "look good" for a job for which they were to assume they were applying, the students obtained low Objectivity scores and low Emotional scores. The very low Objectivity scores would indicate to an interviewer that these persons might be "faking," and their Emotional scores should be compared with other persons of low Objectivity scores.

In a like manner, when the students faked to "look bad" for the jobs, they obtained high Objective and high Emotional scores, and these could again be interrelated in interpretation.⁴

The particular questionnaire described here is used by consulting psychologists as an interview aid. It is concluded from this study that the Objectivity key is a valid key for locating "fakers" and for locating extremely frank respondents, and hence contributes to the interview. It permits the use of different norms for the Emotional key, based on the Objectivity score.

Other findings are that college students fake a questionnaire to make themselves appear socially dominant, as a desirable job characteristic, and socially submissive, as an undesirable job characteristic. This relationship was not found on an industrial sample previously reported. College students fake a lack of drive as an undesirable characteristic, but they cannot, or do not, fake a high drive as desirable.

Received August 18, 1949.

⁴ Readers may wonder why faking "good" and faking "poor" do not yield a consistent pattern on the Objectivity key. The nature of the Objectivity key, and the MMPI L-Scale, is such that opposite directions of faking result in shifts to opposite ends on the Objectivity or L-Scales.

References

1. Hathaway, S. R., and McKinley, J. C. *Manual for the Minnesota Multiphasic Personality Inventory*. New York: The Psychological Corporation, 1943.
2. Longstaff, H. P. Falsability of The Strong Interest Blank and The Kuder Preference Record. *J. appl. Psychol.*, 1948, 32, 360-369.
3. Rothe, H. F. Use of an Objectivity Key on a Short Industrial Personality Questionnaire. *J. appl. Psychol.*, 1950, 34, 98-101.
4. Tiffin, J. *Industrial psychology*. New York: Prentice-Hall, Inc., 1947, rev. ed., pp. 170-171.

Getting Your Message Across by Plain Talk

Arthur O. England

Personnel Planning Office, Air Materiel Command, Dayton, Ohio

In 1948, the Personnel Planning Office conducted an employee attitude survey in the Air Materiel Command (AMC). A sampling of our some 80,000 civilian employees revealed some highly interesting facts about the effectiveness of our communications. Slightly more than one-third of those sampled stated they did not know the procedure for submitting a grievance! The significance of those figures became apparent to our top management when they discovered that there were at least three different publications dealing with the subject of grievances. There was an Air Force Regulation, an AMC Regulation and a Civilian Personnel Letter distributed to each employee. Since approximately 70% of the civilians employed by the Air Force work in the Air Materiel Command, this lack of knowledge of personnel procedure seemed worthy of our attention. We began asking ourselves just how many other procedures and policies of management were unheard of or misunderstood by the employees. Inasmuch as the employees were "informed" by our various publications, the difficulty might very likely lie in the language and style used in writing these directives.

Readability of Our Publications

Using the Flesch formula, an analysis was made of the readability of literally hundreds of Air Force Directives, Civil Service Regulations, Technical Orders, Maintenance Handbooks, employee newspapers, and so forth. Briefly, our study showed that the Air Materiel Command was not getting its message across to the employees:

1. More than 90% of our people found it hard to read and understand our directives.
2. Technical Orders and Maintenance Handbooks used too many big words. Uncommon non-technical words hinder the reader in grasping technical ideas.

3. More than 90% of our people found it hard to read and understand articles in our civilian newspapers.

4. Office memoranda were written in the third person. They were filled with trite phrases. It took too long to read them. Also, messages in that style are hard to remember.

5. More than 60% of our airmen found it hard to read and understand directives addressed to them.

6. More than 20% of our Air Force officers found it hard to read and understand messages addressed to them.

Table 1 shows the "average" reading score of the different publications sampled. Nine different employee newspapers from our field installations throughout the country were sampled.

Obviously, to be able to relate the reading ease scores of printed material to any reading audience, it is necessary to know the educational background of that audience. A study of the personnel records of our civilian employees showed that their educational training very closely approximated that shown for the U. S. adult population. Thus, the U. S. census figures used by Flesch in *The Art of Readable*

Table 1
Reading Ease Survey of Different Types
of Publications

Type of Publication	Average Reading Ease Score	Description of Style
AMC Technical Orders	46	Difficult
AMC Regulations	12	Very difficult
AMC Letters	28	Very difficult
AMC "Daily Bulletins"	25	Very difficult
Hq Office Instructions	13	Very difficult
Air Force Regulations	28	Very difficult
Air Force Letters	17	Very difficult
Civil Service Regulations	28	Very difficult
AMC employee newspapers	45	Difficult

Table 2

Cumulative Percentage of Airmen and of Officers Having Various Amounts of Education

Educational Level	Per Cent of Airmen	Per Cent of Officers
Some grammar school	99.87	100.00
Grammar school graduate	91.97	99.97
Some high school	78.27	99.13
High school graduate	43.47	95.64
Some college	6.39	62.22
College graduate	.42	24.67
Post graduate	.07	7.13

Writing were adapted for use in studying our civilian audience.

Estimates of the educational levels of our military readers are shown in Table 2.

Translating Research Findings into Action

Convinced that the gobble-de-gook in government writing was a real barrier in management's effort to get its message across to the employees, positive steps were taken to sell operating officials on the merits of plain talk. It was recognized that, without the complete support of top management, our re-educational program stood little chance of succeeding. Through the support of Major General J. M. Bevans, Chief of the Personnel and Administration Department, the writer presented a talk before top staff officers of AMC. By using visual aids in the nature of a large, illustrated flip chart, the failure of our gobble-de-gook was shown. Top management enthusiastically indorsed plain talk.

For the next three months, lectures were given on plain talk. Why it should be used, how to use the Flesch formula, and what plain talk would do was discussed. All major components and divisions of the headquarters were covered in this "educational" campaign. At the same time, a work book was prepared showing how to apply the reading ease formula with samples of re-writes of our various publications. After each lecture, these work books were distributed to the audience. Over 2,000 top officials were covered in these lectures. The basic ground work was laid for getting plain talk accepted by management.

The next step meant overcoming "the re-

sistance to change" of thousands of our lower level supervisors. It was decided to publish an official manual to assist all those individuals who do any writing for the Air Materiel Command. Attempting to practice what we were preaching, we prepared the AMC Manual 11-1, entitled, *Gobble-de-gook or Plain Talk?* This manual has a reading ease score of 72; style, fairly easy; audience level, 6th grade. Next, in an effort to attract and hold readership, the manual was illustrated with cartoons. Bold headings typical of commercial advertising were also used on each new topic under discussion. Further, we felt that if this technique of writing our message in an easy-to-read style were to be accepted, some convincing selling had to be done. For many years the federal government and the Armed Forces have been using the same trite, unimaginative, difficult-to-read style of writing. Overcoming that habit pattern would not be easy. The selling points used in the manual were as follows:

1. *Gobble-de-gook is costly.* We had a selling point that is not always applicable to private industry. Every work published is read during the work day. This even applies to our house organs (the civilian newspapers). The point was made that readable language saves reader time. And time saved means money saved for management. It costs \$24,160.00 if our employees spend only ten minutes of their working time reading a four-page directive. This is a modest estimate. It takes ten minutes to read and understand even one page of some of our gobble-de-gook writing. Obviously, if the message in the publication is not understood, then our national defense money is not being spent wisely.

2. *Plain talk saves reader time.* Why not honor your reader's time? Think of the reams of paper work that flow over the reader's desk daily. The writers of these papers are in competition with each other for the reader's attention. Writing stripped of all gobble-de-gook stands a good chance of being read first. It's brief and to the point.

3. *Plain talk style pleases your readers.* Long, wordy sentences confuse the reader. Imagine trying to understand the ideas in one sentence of 461 words. Such a sentence was found in one of our military publications. The reason "Time" and "Reader's Digest" are so popular

is because they please the reader. The average sentence in these magazines has only 17 words. They sell because people want to read them. You have ideas to sell, too. Do you write in a style that pleases your reading public?

4. *Plain talk principles help get ideas across.* It's easy to get into the habit of writing for a mythical audience. But you are writing for real persons to read and understand. They may be busy commanders and division chiefs. They may be supervisors who are more interested in getting out production than wading through a stack of publications. They may be ungraded workers who quit school at the seventh grade. In every case, you must decide who makes up your audience. Who will read your message? If you know the educational background and reading ability of your audience, you will be able to write so they can understand you.

Examples of Using Plain Talk

Here's what a current Air Force directive says about the grievance procedure:

"(1) An employee who has a grievance or his representative will normally present the grievance, in the first instance, orally to the immediate supervisor. The supervisor will consider it promptly and impartially, collecting the necessary facts and reaching a decision. If the employee is not satisfied with the solution of the problem, he will be advised that he may discuss the problem with the next higher supervisor.

"(2) If the employee feels that an interview with the immediate supervisor would be unsatisfactory, he or his representative may, in the first instance, present his grievance to the next supervisor in line. Where an employee feels an interview with the second supervisor would likewise be unsatisfactory he may seek counsel from the civilian personnel officer or his employee relations counselor, whose role will be to advise and aid him in facilitating the employee's approach to a supervisory level determined appropriate by the facts in the particular case."

(Reading ease score, 24; style, very difficult; audience level, college graduate)

Here's how that section of the directive could have been written for the employees:

"Is something about your job bothering you?"

"Here are the steps you can take to solve your problem. In most cases it will be solved at the first step. If not, you have the right to keep going on up to the top. You may present your own case or have someone do it for you.

"Talk with your supervisor. He has been told to give a prompt and fair answer to all problems. Usually, a short, friendly talk with him will fix things up. Be honest and sincere when you talk with him.

"If you feel that your supervisor will not handle your case fairly, you may go directly to his supervisor. Or, if you have gone to your supervisor and he didn't handle your problem to suit you, you may still go to his supervisor.

"If you feel your case has not yet been, or will not be, handled fairly by either of them, go to your personnel technician. He can't give you a final answer, but he can tell you how to get it."

(Reading ease score, 85; style, easy; audience level, 5th grade)

Tips on Writing to Be Understood

In addition to presenting an explanation and examples of using the Flesch formula with potential reader audience, the following tips on writing to be understood were presented in our manual:

The key question to all writing is, "What am I trying to tell *whom*?"

If you consider the following points, you may be fairly certain your message will be understood.

Define your audience. When you sit down to write, the first thought that should come to your mind is, "Who will read it?" As a rule, material that is written for the base commander with college training is not suitable for the ungraded employee who quit school at the eighth grade. Writing can be simple enough to be read with ease and understanding by a poor reader and yet be interesting enough to hold the attention of a good reader.

Define your purpose. Just what are you trying to say? Is your purpose to get employees to save their sick leave? Or is it to explain the benefits of sick leave? If you are

not clear in your mind about what you want to say, it's a sure bet your reader won't know either. The purpose for writing is of foremost importance. The reader should be able to understand what to do, why it must be done, and how to do it.

Present your ideas in logical order. A simple easy flow of related ideas is necessary if your message is to get the effect you want. Each part of your message should prepare the reader for what is to come. Don't jump from one idea to another. Complete your discussion of each idea before introducing another. Present one idea at a time.

Avoid unnecessary technical "niceties." Don't use fine distinctions in words when they are not needed. Writers often spend too much time quibbling about technical niceties which have no real meaning for the reader.

Keep the vocabulary familiar. Logic will not help the reader if he does not understand the words used. In AMC our biggest job is to find ways of writing about technical ideas. These ideas are often complicated and some technical words must be used. But the words used to modify these technical words should be familiar ones. Non-technical terms that will cause trouble for the reader should be omitted wherever possible. A technical idea is hard enough to grasp without also including hard, non-technical words to confuse the reader.

Use simple sentences. It was the fashion many years ago to write articles with sentences running well over 100 words. Today edu-

cators clearly show us that writing can be more easily read and remembered if the sentences are short. In order to do this, you should avoid using involved sentence structure. Involved sentences are too much of a mental burden for most readers. Why ask your reader to expend mental effort trying to figure out what you are trying to say? Be brief and honor the reader by telling facts in short, brisk sentences.

Use words of one or two syllables. Avoid words that will stop the reader. Use the good short ones that come first to your mind. Edit your paper from your reader's point of view before signing it. These few words found in current AMC directives are typical of the "stoppers" we mean:

recapitulation	idiosyncracies
beleagured	adjudication

"Stoppers" add to the reader's difficulty in getting the message. Foreign phrases should be avoided. Also, don't use short words that are not common. Using words of one or two syllables is not the whole answer. The words must also be understood.

There have been many noticeable improvements in our communications since the inception of the campaign for more plain talk in government writing. But like all ingrained habits, it will take time to unlearn the old ones and adopt the new ones.

Received March 13, 1950.

Early publication.

Prediction of Academic Success in Three Schools of Nursing

Albert H. Ford

Towson, Maryland

There are various reasons why nurse trainees withdraw from schools of nursing; paramount among them is academic failure. In a study by Horner (1), reported by Potts (3), approximately 37 per cent of a group of more than 15,000 students admitted to schools of nursing over a period of years were eliminated prior to completing their courses. Of those eliminated in two years' classes, about 30 per cent withdrew because of academic failure. Potts (3) also found classroom failure the largest single reason for withdrawals from schools of nursing.

Perhaps more significant than the total percentage of withdrawals from schools of nursing is the percentage of withdrawals in the early stages of training. Horner (1) found that 63 per cent of those withdrawing did so in the preliminary period of training; 84 per cent had left the school by the end of the first year. Potts (2) points out that in a particular group of 1,555, approximately 90 per cent of the eliminations for classroom failure came within the first six months of their course.

It can be concluded from the foregoing that academic failure is one of the principal reasons for withdrawal from schools of nursing, and further, that failures occur early in the school's program. If we assume that better selection techniques are able to predict in advance an applicant's likelihood of academic success, then these techniques will save much of the expense ordinarily incurred by the withdrawal of unsuccessful trainees.

It was for reasons such as these that the nurse training supervisors of three Knoxville, Tennessee, training schools decided to solicit the aid of the University of Tennessee in the development of a more efficient selection program.

Statement of the Problem

Although the general problem involved was twofold—the determination of scholastic success, and also success in ward training or on-the-job training—this study was concerned only with the former phase. Its specific

purpose was to determine the extent to which scholastic success in Knoxville hospital training programs could be predicted from a battery of tests administered to entering groups.

Subjects

The subjects used in this study were trainees accepted in the training programs of the East Tennessee Baptist Hospital, the Fort Sanders Hospital, and the Knoxville General Hospital. The original sample consisted of 187 trainees from six groups admitted to the schools without regard to the scores made on the selection tests. Each hospital contributed from one to three groups. With the exception of one group, all trainees were tested within 30 days of their acceptance into training. The other group had completed from one-half to one year of its particular program.

The school entrance requirements are such that applicants must be females between the ages of 17 and 35. They must have graduated from an accredited high school with 16 units of credit and be eligible for admission to the University of Tennessee.

The Variables

The original battery of predictors included the following measures:

1. George Washington University Series Reading Comprehension Test for Prospective Nurses; Form 1, First Edition; by Thelma Hunt.
2. George Washington University Series Arithmetic Test for Prospective Nurses; Form 1, First Edition; 1940; by Thelma Hunt.
3. American Council on Education Cooperative General Science Test; Revised Series, Form X; by Paul E. Kambly.
4. Science Research Associates Primary Mental Abilities; 1948; by L. L. Thurstone and Thelma Gwinn Thurstone.

5. Kuder Preference Record; Science Research Associates; Form BB, 1942.
6. High School Point Average.

The criterion for the study was the point average of all scholastic grades earned by the trainee up to March, 1949, either at the hospital or on the University of Tennessee campus. If the trainee left the hospital prior to that time, her average was based on all courses completed prior to withdrawing. Each trainee took all of the tests; high school averages were available for all trainees except one who was admitted on the basis of Veterans Administration high school proficiency examinations.

Procedure

Since the number of cases available ($N = 187$) did not justify the inclusion of all the separate measures, the most promising in terms of predictive power were sought on the basis of the validity coefficients computed for the first two groups. From these correlations it seemed likely that the best measures to include in the final battery were: Reading Comprehension Test; Arithmetic Test; Cooperative General Science; Total Primary Mental Abilities; High School Point Average; and the Science and Social Service scales of the Kuder Preference Record.

The validity coefficients of the various Kuder scales were both small and inconsistent. The largest obtained was .34 between the Scientific Scale and the scholastic point averages of the group; however, the second group revealed a coefficient of only .005. The Scientific and Social Service scales were retained in the final battery more on the basis of reports in the literature and the significant group centile ratings on the measures than for their validity coefficients. From the validity coefficients of the subtests of the Primary Mental Abilities, it did not appear that there was any additional predictive power to be gained beyond that provided by the total of that measure.

Similarly, in terms of the total sample available, it was considered advisable to combine the various groups from the three hospitals, thus considering them as derived from a homogeneous population. To ascertain if any differences did exist between the groups which would preclude treating them collectively

rather than as separate hospital groups, an analysis of variance of the differences between groups was made on the variables to be included in the final battery.

The F-ratios revealed that very significant differences existed on the Reading Comprehension and Kuder Social Service, differences which were not likely to occur by chance alone one time in a hundred. The F-ratio found on the reading test was most striking and was the impetus for further analysis to determine the particular areas of greater difference.

As was pointed out earlier, the students comprising one group had been in training from one-half to one year prior to taking the battery of tests, whereas the other groups were tested within 30 days of their admittance to the program. By observing the F-ratios depicting the within-group and between-group variances of all six groups, it became apparent that half of the total between-group variance was being contributed by the trained group on the reading test as well as on the arithmetic test. It seemed likely that by omitting the trained group, much of the variance in the reading test as represented by an F-ratio of 5.31 would be eliminated.

With the trained group eliminated, none of the F-ratios except the one for the scientific scale of the Kuder Preference Record were significant at the 1 per cent level of confidence with 4 and 141 degrees of freedom. Four of the measures, including the reading test, did not reveal differences significant at the 5 per cent level of confidence. The Cooperative General Science Test and the Social Service scale of the Kuder Preference Record revealed differences at the 5 per cent level, but not at the 1 per cent level. The F-ratio for the science test was 2.57, with a ratio of 2.44 being significant at the 5 per cent level with 4 and 141 degrees of freedom.

From the above, it was concluded that, with the exception of the Kuder Scientific scale, which was later eliminated from the battery, the differences which existed were not of a magnitude which would preclude combining the groups from different hospitals, thus regarding them as derived from a homogeneous population. When we consider that all of the hospitals have similar academic programs, that their requirements for entrance are similar,

Table 1

Criterion Correlations and Intercorrelations between Variables and the Mean and Standard Deviation of Each Variable

Note: N = 137 except in variable High School Point Average where N = 136.

Test	Read	Arith	Scn	PMA	Scn(K)	S.S(K)	HSPA	HTPA	Mean	S.D.
1. Reading Comprehension		.54	.56	.56	-.13	.11	.38	.56	62.9	13.4
2. Arithmetic			.40	.57	-.02	.04	.33	.39	24.9	9.8
3. A.C.E. Science				.47	-.12	-.02	.33	.57	26.8	9.4
4. Total P.M.A.					-.14	-.03	.46	.48	158.5	36.7
5. Scientific (Kuder)						.14	-.01	.00	60.4	13.2
6. Social Service (Kuder)							.07	.01	102.3	13.0
7. High School Point Average								.51	2.4	.6
8. Hospital Training Pt. Average									2.0	.9

and further, that they admit applicants from approximately the same area, it seems reasonable that such influences would assure that the applicants were derived from a reasonably homogeneous population.

Thus, after eliminating the trained group from the over-all sample, there remained five groups with a total of 146 individuals.

The next phase of the study consisted in determining the criterion correlations and the intercorrelations of each of the final measures of the five untrained groups.

Table 2

Table of Beta Weights and Coefficients of Multiple Correlation Showing Successive Changes as Measures Are Eliminated from the Battery

Test and Battery	Beta Weights	Multiple R
Coop. Science	.304	
Reading Comp.	.242	
High Sch. Pt. Av.	.286	.699
Total P.M.A.	.060	
Arithmetic	.014	
Coop. Science	.304	
Reading Comp.	.246	
High Sch. Pt. Av.	.286	.699
Total P.M.A.	.056	
Coop. Science	.312	
Reading Comp.	.254	
High Sch. Pt. Av.	.299	.698
Arithmetic	.034	
Coop. Science	.316	
Reading Comp.	.269	.697
High Sch. Pt. Av.	.303	

Table 1 gives the product moment correlations between the variables, with the mean and standard deviation of each variable. Both scales of the Preference Record yielded very low validity correlations; in fact, these two measures correlated neither with themselves nor with any of the other measures to an extent greater than .14. It was at this point that the Scientific and Social Service scales were dropped from the battery.

Table 2 gives the beta weights and multiple correlation coefficients for various combinations of the final measures, showing successive changes as measures with lower criterion correlations are eliminated from the battery.

It was concluded from Table 2 that the most practical regression equation for the prediction of scholastic point averages for students from the general population concerned in this study would include the reading test, the science test and the high school point average, which gives a multiple R of .697 as compared to an R of .699 for all measures. The other measures, although giving criterion correlations as high as .48 (Table 1), do not add sufficiently to the predictive power of the battery to justify the time and expense involved in administering and scoring the tests.

Shrinkage of the Multiple R

When the coefficient of multiple correlation is determined from a given set of data, as above, and is applied to a second set of data, the yield in the latter case will be less, even though the second set of data is strictly comparable. This shrinkage of the multiple R

varies with the number of variables contained in the regression equation, the number of cases, and the size of the coefficient of correlation. A shrinkage-deduction formula has been devised by Smith (4) to apply to the coefficient of multiple correlation which provides a more accurate estimate of the multiple R. When this formula is applied to the present data a multiple R of .686 is forthcoming, indicating that shrinkage to the extent of .011 can be expected when the regression equation is applied to a second group.

The formula for the most economical prediction of the criterion in terms of raw scores was:

$$\text{Hospital Training School Average} \\ = .017 X_{\text{Read.}} + .029 X_{\text{Sc.}} + .420 X_{\text{H.S.P.A.}} - .87.$$

Conclusions

On the basis of the study the following conclusions seem to be justified:

1. The science test, reading test and high school point average were fairly effective in predicting success in the schools of nursing, the multiple correlation coefficient being .697.

2. Although the arithmetic test and Total P.M.A. correlated rather well with the criterion, .392 and .479 respectively, neither of these tests, either separately or together, added appreciably to the predictive power of the battery. Increases in the multiple R were significant only in the third decimal place.

3. The sub-tests of the P.M.A., in general, correlated less highly with the criterion than did the total of that measure, as evidenced by validity coefficients computed for two early groups.

4. Although the averages of the total group showed its members to be more interested in

science and social service, as measured by those two scales of the Kuder Preference Record, than women in general, neither scale contributed anything to the predictive power of the battery. The validity coefficients were as follows: Scientific, .000; Social Service, .010.

5. Significant differences between groups from different hospitals appear when one group has had one-half to one year's training prior to taking the tests and the remaining groups are tested within 30 days of their acceptance into the hospital's training program. Such differences were most striking on the Reading Comprehension test; the F-ratio for that test being 5.31 with the trained group included in the total population. Although such an F-ratio indicates differences between the groups which are significant at the 1 per cent level of confidence, with the trained group eliminated, existing differences are not significant at the 1 per cent or 5 per cent levels of confidence.

6. Had the regression equation forthcoming from this study been used in selecting the trainees participating in the study, 13.1 per cent of them would not have been admitted (assuming 1.50 as the hospital critical score and the score one P.E. below 1.50, or 1.08 as the cutting score).

Received August 25, 1949.

References

1. Horner, H. H. *Nursing education and practice in New York State with suggested remedial measures*. Albany: University of the State of New York, 1934, pp. 38.
2. Potts, Edith Margaret. The selection of student nurses. *Amer. J. Nurs.*, 1941, 41.
3. Potts, Edith Margaret. Use of tests in selecting student nurses advantageous to hospital and student. *Hospital Mgmt.*, 1941.
4. Smith, B. B. Forecasting the acreage of cotton. *J. Amer. statist. Ass.*, 1925, 20, 31-47.

Critical Requirements for Dentists *

Ralph F. Wagner

American Institute for Research and Department of Psychology, University of Pittsburgh

An attempt to improve selection methods at the School of Dentistry led to the discovery that no systematic investigation had been carried out to determine the characteristics either of a successful dental student or an effective practicing dentist. Since this information is important for the development of personnel procedures, research was undertaken at the University of Pittsburgh to obtain a precise and practical definition of requirements for the profession.

The method employed is called the critical incident technique. Information on the procedure has only recently appeared in the literature.¹ It was developed in order that a comprehensive list of behaviors of the kind which make the difference between success and failure in an activity or profession might be obtained. Persons in or normally associated with the profession and who are considered qualified to judge competency with respect to one or more phases of the job are asked to describe the most recent incident they observed in which a participant carried out a part of the job either in a particularly effective or ineffective manner. They are asked to describe the situation, the relevant circumstances and exactly what the participant was observed to do. No interpretation is requested regarding abilities, aptitudes, motivations, and attitudes, which might have been responsible for the behavior.

After a large number of incidents are obtained, an analysis is made to determine the specific behavior which caused the observer to judge an individual as effective or ineffective. A grouping and structuring of these behaviors result in: (1) a list of those aspects of the job which are "critical" in the sense that they

caused an observer to make a judgment of job effectiveness; and (2) a series of specific statements of the contrasting ways in which effective and ineffective job participants behave in carrying out these aspects. The "critical requirements" of the occupation are then derived through an analysis and a summarization of these contrasting ways of carrying out each aspect.

The Present Study

Incidents were obtained from three sources: (1) patients; (2) dentists themselves; and (3) instructors in dental school clinics. Patients were not expected to supply data regarding the technical aspects of dentistry; this information was to be obtained from dentists and instructors. As expected incidents supplied by patients dealt more with personality, manner, business practices, appearance of office, and similar factors. Clinic instructors provided particularly useful data for determining requirements both for success in dental school and for effectiveness in general practice. An instructor has an opportunity to observe dental practice being carried on by persons often varying considerably in skill and is in a position to know the important details when behavior which is critical occurs. They correspond to supervisors who have, in other studies, proved to be a source of particularly useful data. In the case of practicing dentists there is no individual in a position comparable to that of an instructor or supervisor. In most cases only the dentist who was doing the actual work is acquainted with all relevant circumstances: Dentists were therefore requested to describe incidents concerning their own practice.

In securing incidents from all three sources—patients, dentists, and instructors—either the word "effective" or "competent" was used in referring to the desirable type of dentist. These words were purposely used in place of "successful" because of the monetary connotation of the latter. Moreover, it was requested

* The study was conducted in close cooperation with Dr. L. E. Van Kirk and Dr. W. F. Swanson, School of Dentistry, and Dr. J. C. Flanagan, Department of Psychology. Dr. Wagner, who carried out the research, is Project Director with the American Institute for Research and Lecturer in the Department of Psychology. ¹ Flanagan, J. C. Job Requirements, From Dennis, W. (Ed.), *Current Trends in Industrial Psychology*. Pittsburgh, University of Pittsburgh Press, 1949.

that the identity of the individual whose performance had been observed not be given either in effective or in ineffective incidents.

Patients were requested to describe two kinds of incidents—those in which the dentist's performance had caused them to recommend him enthusiastically to a friend and those in which his performance had caused them to change or consider changing to a new dentist. Dentists were also asked to describe such incidents in which it had been their own performance which was responsible for the patient's action. In addition, however, they were asked for incidents in which the patient was unaware that particularly effective or ineffective performance had occurred but which nevertheless had caused the dentist himself either to feel a great deal of professional satisfaction or to feel that he would perform more effectively if given a second opportunity. Clinic instructors were asked to describe incidents in which they had observed a student perform in either a particularly effective or ineffective manner. Particularly effective performance was defined as performance the instructor might wish to cite in the classroom, insist that all students copy, or the kind which would contribute significantly to the student's effectiveness if he were in practice. Ineffective performance was defined as the kind which, if it occurred repeatedly, or even once under certain circumstances, would cause the instructor to doubt seriously the student's probable effectiveness in practice.

A total of 781 incidents were obtained,—257 from patients, 359 from dentists, and 165 from clinic instructors. One of the most interesting findings of the study was the enthusiastic manner in which dentists participated. There was some concern at first as to how willing an individual would be to describe ineffective incidents concerning his own performance. It was found, however, that dentists gave such incidents as fully as they gave effective incidents. The following example is typical.

About six months ago a boy of sixteen for whom I had done several fillings came in for his regular weekly appointment. I was rushed and did a rapid filling for him, not being careful about the depth of cavity preparation. No cement base was placed. Patient returned several days later complaining of severe toothache. I removed filling and found pulp ex-

posure. Refilled tooth with sedative cement. Patient did not return but a dentist friend said he extracted the tooth for the patient several weeks later.

Another example will further illustrate the nature of the data collected in the study. This incident was obtained from a clinic instructor and should be of interest to dental educators.

Student came to me with a gold inlay filling. He told me filling was not good and he would like to do the work over. On examining the filling in the mouth I found the filling was not quite as bad as the picture he had given me and could very possibly have been made satisfactory. However this student showed both the knowledge and was conscientiously interested in the patient's welfare and his own work to repeat the work. Most students are satisfied with only fair work and do not generally want to repeat it to make it better.

An analysis of the incidents from all three sources indicated that there were four main aspects in serving as a general practitioner. The following titles seemed descriptive of these aspects: I. Demonstrating Technical Proficiency; II. Handling Patient Relationships; III. Accepting Professional Responsibility; and IV. Accepting Personal Responsibility.

The critical behaviors under each of the above main areas were grouped into sub-areas. The behaviors under Area I were grouped according to the specific treatment being rendered. Development of sub-areas within the other major areas, however, was accomplished on logical grounds, guided by the actual nature and distribution of the behaviors. Although the behaviors in these areas were not related to any specific type of treatment, it was found that they could be grouped into relatively discrete sub-areas.

As a means of summarizing the content of the incidents, definitions of major areas and sub-areas were written. These definitions provided a detailed description of the nature of the dentist's work and responsibilities as derived from the 781 incidents analyzed in the study. As a second means of summarizing the content of the incidents, a tentative group of 40 "critical requirements" were defined. They consist of a series of statements which express the specific way an outstanding dentist performs in the important situations which are characteristic of his profession. Some indi-

Table 1
Distribution of Critical Behaviors Among the Four Major Categories

Source	Technical Proficiency	Patient Relationship	Professional Responsibility	Personal Responsibility	Total
Patients	107	168	6	94	375
Dentists	138	160	37	95	430
Instructors	65	36	15	59	175

cation of the relative importance among the 40 critical requirements is provided by the frequency of their occurrence in the incidents from each of the three sources.²

Although it is impossible to present the 40 critical requirements and their frequencies in the present paper, the number of critical behaviors falling into each of the four major categories, broken down according to the source of the incident, is shown in Table 1. These frequencies exceed the number of incidents collected in the study since an incident often contained more than one critical behavior.

Summary

The present study furnished information regarding the critical aspects of dental practice. Its purpose was *not* to produce a curriculum for training dental students. The study does, however, furnish information on the specific kinds of dental practice which have frequently made the difference between effectiveness and ineffectiveness both in practice and in dental school clinics. The results indicate the areas which *most frequently* cause difficulty while the student is in the clinic and after he has gone into practice and provide information on the type of practice which the patient, the dentist, and the clinic instructor consider particularly effective. The conclusions which these results suggest are as follows:

1. The requirements for effectiveness in dentistry are complex. They are not confined alone to the demonstration of technical

² Persons interested in a fuller description of the study, including the tentative statements of critical requirements and definitions of areas and sub-areas, should write to the School of Dentistry, University of Pittsburgh, Pittsburgh 13, Pa. Also, this information may be ordered as Document No. 2826 from American Documentation Institute, 1719 N Street, N.W., Washington 6, D. C., remitting \$.50 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$1.10 for photocopies (6 × 8 inches) readable without optical aid.

proficiency. Although the critical nature of this area is strongly supported, there are non-technical behaviors which are also critical to effectiveness. "Handling Patient Relationships" is a particularly important area. Other critical areas, to use the titles adopted in the present research, are "Accepting Professional Responsibility" and "Accepting Personal Responsibility."

2. Many of the characteristics which have commonly been accepted as important for effectiveness in dentistry must be re-evaluated. "Ability to converse on topics of the day," for example, has been mentioned as important upon various occasions yet in the 781 incidents analyzed in the present study it appeared only once. The same was found to be true of "asking question when the patient is unable to answer." And, although applicants have in the past been rated upon voice quality, this was not found to be a factor in any incidents. In comparison, "discussing treatment being planned or rendered" occurred in 61 incidents. In view of the complexity of requirements which the present study suggests, it would seem advantageous to concentrate on those which are critical.

3. Critical behaviors revealed by the present study provide an additional basis for evaluating clinic performance. A form, developed in cooperation with persons experienced in dental education, for recording systematically observed occurrences of critical behavior would provide objective evidence on which to base judgments of effectiveness. These judgments would be closely related to the requirements of actual practice. As such they would reflect the adequacy of training, indicate eligibility for graduation, and provide the realistic criterion which selection batteries might strive to predict.

Received March 22, 1950.

Early publication.

The Intra-Individual Relationship Between Interest and Ability

S. M. Wesley, Douglas Q. Corey, and Barbara M. Stewart

University of Southern California

An understanding of problems related to educational and vocational guidance has become increasingly important during the post-war years. Advisement of veterans, by arrangement with the Veterans Administration, alone has resulted in the setting up of centers which annually provide this service to many thousands of former servicemen.

A question which has long been considered a primary concern in this field is that of the relationship between vocational interests and vocational abilities. Summaries of studies concerned with the problem are presented by Strong (4) whose review illustrates a trend in the direction of more adequate methodological procedures for determining the degree of this relationship. Early investigations compared various interests on a vocational interest test to overall ability as measured by a single criterion such as intelligence test scores or college grade point averages. The results of such studies, in general, indicated low or negligible correlations. An explanation for this seemed to lie in the fact that different interests were matched with a single general ability rather than with specific abilities corresponding to those interests.

When studies were designed in which each interest was matched to a corresponding ability, higher correlations were obtained. A profitable approach (resulting in significant positive correlations ranging from .32 to .40) was that of Triggs (5) who compared interest scores of one hundred college men on the *Kuder Preference Record* to corresponding ability scores on the *Iowa High School Content Examination*. However, the magnitude of the relationship shown by such studies continued to be surprisingly low.

Since the ability scores used in Triggs' study were based on deviations from the group means, the following question suggests itself: What is the relationship between interest and ability when scores for each individual represent deviations from his own mean, rather than from the group mean?

Some evidence that this approach might result in higher correlations is given in the study of Segel (2). He compared interests, as measured by the *Strong Vocational Interest Blank*, with abilities as measured by the *Iowa High School Content Examination*. However, his method was unique in that he correlated interest scores with *differences* between two ability scores, and obtained higher correlations than those found between an interest and an *absolute* ability score in a corresponding area. For example, scores on the Engineer key correlated .57 with the difference between scores in Mathematics and scores on a History and Social Science test, while the same interest correlated only .49 with scores in Mathematics alone. Although the differences between these two types of correlation were not statistically significant (results were based on one hundred cases), they do suggest that the use of more than two abilities in relation to each other might point the way to even higher correlations than those previously shown. The writers felt that if a truly relative score were derived by first calculating a mean ability score for each individual, a greater relationship might be shown to exist.

Procedure

In order to investigate this hypothesis a study (1, 3) was conducted in which tests of interest and ability were administered to 156 male college students enrolled in an introductory psychology course. Since tests were administered on different days the number taking any one test was not constant and ranged from 115 to 132. The measure of interest used was the *Kuder Preference Record* which yielded scores in the following areas: mechanical, computational, scientific, artistic, literary, musical and clerical.¹ Measures of ability were selected to correspond to these

¹ The persuasive and social service categories were not included because of the lack of adequate tests of ability in these areas.

interest categories as follows: Survey of Mechanical Insight; Stanford Arithmetic Test; Iowa High School Content Examination, Section 3, Science; The Meier Art Judgment Test; Iowa High School Content Examination, Section 1, English and Literature; Seashore Measures of Musical Talents, Series A; and Minnesota Vocational Test for Clerical Workers.

Total scores were computed for all tests and were equated by converting to standard scores. Pearson r correlations were first obtained between interest and its corresponding ability in each of the seven areas. This was done in order to provide data based on the traditional, or inter-individual method, to which results of the new, or intra-individual method might be compared.

To obtain correlations based on scores representing deviations from the individual's own mean, rather than that of the group, the mean ability score for each subject was computed (in terms of standard score units). The differences between his separate ability scores and his own mean level of ability were next determined. These differences then became the measures of relative ability to be correlated with the corresponding measures of relative interest. Since the *Kuder Preference Record* is so constructed that every item is chosen at the expense of another, it was assumed that the resulting interest scores were approximately relative to the individual mean and that further manipulation of these data was therefore not necessary. Pearson r correlations between *relative* interest and *relative* ability, for each of the seven areas, were then computed.

Since these correlations indicated the relationship only for the group as a whole, it was considered desirable to investigate as well the variation in interest-ability agreement for individuals within this group. Accordingly, for each of the one hundred students who had completed all tests and for whom complete data were therefore available, the standard scores for his seven areas of interest and ability were arranged in two rank-order sequences. Rho coefficients of correlation were then computed for each of the individual sets of scores.

Results

There were thus obtained, first, seven interest-ability correlations based on deviations from group means, and seven correlations based on deviations from individual means. As may be seen from Table 1, these ranged from .07 to .47 (not corrected for attenuation) where group means were used, and from .23 to .68 where individual means were used. The seven "group mean" correlations were averaged (after transforming to Fisher z values) and resulted in a mean Pearson r correlation of .30. The average of the seven "individual mean" correlations was .42. The t ratio, based on the difference between mean z values, was 3.3 which is significant above the 1 per cent level of confidence.²

The 100 individual rank-order correlations ranged from $-.57$ to $+1.00$, and (by transforming to Fisher z values) a mean of $+.46$ was obtained. The difference between this mean and that of .30 for the "group mean" correlations also resulted in a t ratio of 3.3, significant above the 1 per cent level of confidence.

It was thus shown that significantly higher correlations are obtained by the use of the two methods presented here than by the more traditional method of comparing interests and abilities. It is to be expected that with the development of better tests, and with more exact matching of tests of interest and ability, an overall relationship in excess of that found here may be demonstrated. Although in every vocational area the correlation was increased when individual levels of interest and ability were used, in calculating the mean correlation substantial relationships in some areas were offset by low relationships in others. This variation in the correlations obtained may well be, in part at least, a function of variation in the extent to which each of the ability tests reflects experience as opposed to aptitude. It would be expected that a test which is more heavily weighted with the experience factor would show a higher relationship with interest

² This difference is even more significant when it is considered that the formulas available for testing significance of differences between correlations assume that the figures are obtained from independent random samples. In this study, of course, the correlations were obtained from the same sample, so that any such test of significance would err on the conservative side.

Table 1

Correlations Between Interest and Ability Based
on Deviations from Group and from
Individual Means

Vocational Area	Group Means		Individual Means	
	N*	r	N*	r
Mechanical	131	.44	126	.50
Computational	115	.24	112	.47
Scientific	126	.33	126	.35
Artistic	131	.29	127	.31
Literary	125	.47	125	.68
Musical	122	.21	118	.23
Clerical	132	.07	125	.33
Mean		.30		.42

* The size of N in each case depended on the number of subjects for whom necessary data were available.

because of the correlation which we know to exist between interest and experience. An examination of the correlations in Table 1 shows that those tests in which the experience factor would play a larger part, such as the Iowa Literary Test and the Survey of Mechanical Insight, have a higher correlation with interests than do those ability tests such as the Minnesota Clerical and the Seashore Tests which have a lower weighting with the experience factor. However, there still exists the reasonable hypothesis that there is a genuine variation in the degree of relationship between interests and abilities for different activity or vocational areas.

The Meaning of Individual Differences in Interest-Ability Congruency

The wide range in interest-ability congruency for different individuals ($-.57$ to $+1.00$) raises what may be an important question for those concerned with vocational guidance: Why for some individuals is the correlation high and positive, while for others it is low or even negative? In order to explore this area, age, intelligence and personality factors were studied to determine whether they might be related to these individual differences. The measures of intelligence and personality, respectively, were the *Army Alpha Examination*, *First Nebraska Revision*, and the *Minnesota Multiphasic Personality Inventory*. An upper

and lower 25 per cent of the group were selected, based on those having highest and those having lowest interest-ability congruency as shown by the individual rank-order correlations. For these extreme groups a comparison was made of mean age and of mean intelligence scores, but the differences were found to be insignificant. Similar comparisons were made between mean scores obtained by the upper and lower groups for each of the nine categories of the *Minnesota Multiphasic Personality Inventory*. Although only one significant difference (for the Schizophrenia scale) was found, the group having highest agreement between interests and abilities showed scores on eight of these scales above those of the other group, and in a direction away from the level of "normal" adjustment. The meaning of this finding is not clear, but it appears to be related to a greater differentiation in interests and in abilities for those having a less adequate personality adjustment.

It is felt that a different test of personality, concerned with basic character structure rather than with nosological groups, might reveal important differences between those individuals whose interests and abilities are in agreement and those where marked deviations are found. It is possible that the use of projective measures of personality would demonstrate such a relationship and contribute to an increased understanding of these individual differences in interest-ability congruency.

Predictability of Extreme Ranks

In computing the rank-order correlations between interest and ability for the individual subjects in the study, a tendency was noted for high ranking and low ranking ability areas to have greater predictability in terms of interest test scores than those ability areas in the middle of the individual's range.

Further analysis of the data was therefore made to determine the predictability of each interest rank from ability rank. A study of the rank-order sequence of interest and ability for each individual showed that for 31 per cent of the cases the highest interest fell in the same area as the highest ability. For 20 per cent of the cases, the second highest interest was in the same area as the highest ability. Con-

tinuing this procedure, the percentages became increasingly smaller so that for only 4 per cent of the cases was the seventh, or lowest, interest in the same area as the highest ability. Thus, for 51 per cent of the group, the first or second highest interest was in the same area as the highest ability. Chance would have permitted only 29 per cent, and predictability was thus 22 per cent better than chance.

Similar procedures were applied in studying the number of cases in which the second highest interest fell in the same area as the second highest ability, and so on throughout the seven different ranks. It was shown that prediction from ability to interest for rank one and for rank seven was better than prediction for ranks two through six. This is explained in part, of course, by the fact that, at the ends, the error can extend in only one direction, while in the middle it can vary in two directions.

It is probably this end-effect phenomenon which has led vocational counselors to the clinical belief that there is a higher relationship between interests and abilities than has been shown by statistical studies, where the intervening ranks must also be considered and where overall accuracy of prediction is lowered accordingly. In observing generally good agreement between highest interest and highest ability, and between lowest interest and lowest ability, counselors have to this extent at least considered the ordinal position of both, and have thus actually based guidance on the relation of scores to the individual's own mean rather than to that of the group. Although increased accuracy of prediction from extreme ranks may be largely a statistical artifact, the use of such prediction for purposes of vocational guidance would appear to be justified.

Summary

The purpose of this study was to investigate the relationship between vocational interests and abilities when the magnitude of test scores is relative to the individual's own level, rather than to the group level of interest and ability.

The procedure and findings were as follows:

1. The *Kuder Preference Record* and ability tests corresponding to seven of the Kuder interest areas were administered to 156 male college students.

2. A Pearson r correlation was first obtained between each interest and its corresponding ability for scores based on deviations from group means. The mean of these seven correlations was .30.

3. Pearson r correlations between interest and ability were then obtained for scores based on deviations from individual means. The mean of these seven correlations was .42 which was shown to be significantly higher than the mean of .30 derived from the scores based on deviations from group means.

4. Rank-order correlations between the seven interest areas and the seven ability areas were computed for one hundred individuals. The mean of these correlations was .46, which was shown to be significantly higher than the mean Pearson r correlation of .30 derived from scores based on deviations from group means.

5. There was shown to be a wide range of individual differences in interest-ability congruency. Rank-order correlations ranged from $-.57$ to $+1.00$. For the 25 per cent of the group having highest and the 25 per cent having the lowest interest-ability correlations, no significant difference was shown between mean age and mean intelligence. However, the group having the highest agreement did show a tendency to less adequate personal adjustment in that mean scores on eight of the nine scales of the *Minnesota Multiphasic Personality Inventory* were higher than mean scores for the group having lowest interest-ability agreement. Only one of these differences, that for Schizophrenia, was significant.

It was suggested that the use of projective techniques would be a useful approach in further study of the personality factors related to this individual variation in the congruency between interest and ability.

6. For individual interest and ability scores arranged in rank order, it was shown that prediction from the extreme ability ranks, one and seven, to the corresponding interest ranks was much better than prediction from the intervening ranks, two through six. Although this is largely an artifact due to the end-effect, and not a true difference in efficiency of prediction, it has probably led vocational counselors to consider the extreme ordinal positions of interests and abilities for a given indi-

vidual, and to apply these findings in their vocational guidance work. Such a procedure would seem to be justified and of value when applied in this manner.

Received August 31, 1949.

References

1. Corey, D. Q. A comparison of two methods of determining the relationship between vocational interests and abilities. Unpublished Master's thesis, University of Southern California, 1947.
2. Segel, D. Differential prediction of scholastic success. *Sch. and Soc.*, 1934, **39**, 91-96.
3. Stewart, Barbara M. A study of individual variability in the relationship between interest and ability. Unpublished Master's thesis, University of Southern California, 1947.
4. Strong, E. K. *Vocational interests of men and women*. Stanford University, California: Stanford University Press, 1943.
5. Triggs, Frances O. A study of the relation of *Kuder Preference Record* scores to various other measures. *Educ. psychol. Measmt.*, 1943, **3**, 341-354.

A Projective Test for Vocational Research and Guidance at the College Level *

Robert B. Ammons

University of Louisville

Margaret Newman Butler

Colorado Woman's College

and

Sam A. Herzig

University of Denver

In their professional work the clinical psychologist and vocational counselor are often called upon to relate for a given individual problems associated with occupation to underlying personality structure. This need to relate personality and vocations has been discussed at some length by Bixler and Bixler (7), Darley (10), Kilby (14), and Trabue (23). As a practical matter, it is often necessary to recognize stated vocational problems as symptomatic of deeper personality imbalances, and to attempt to treat these more fundamental disturbances prior to attempting to find solutions for the vocational problems.

Berkshire, Bugental, and Cassens (6) in a survey of tests used in guidance centers report that by far the most frequently utilized personality instruments are of the paper and pencil type, the Bell Adjustment Inventory, the Minnesota Multiphasic, and the Bernreuter Personality Inventory. The known clinical inadequacy of these tests is such that no comment is necessary here. The Rorschach is reported as being used in only 20 per cent of the guidance centers in the sample, and the Thematic Apperception Test is not even listed among the 79 tests estimated to be most frequently used. Although formal projective devices are apparently not widely used in guidance centers, they are accepted as being valid for general clinical use, and are sometimes used in conjunction with counseling (2).

* Plates and manual (1) can be obtained from R. B. Ammons. Thanks are due Professor R. A. Irwin of the University of Nevada, Professor R. B. Winn of Monmouth College, and Mrs. Carol Ammons of the University of Louisville for critical reading of this article and many helpful suggestions.

In his comprehensive review of projective techniques, Bell (4) mentions only a few systematic attempts to link projective methods of investigating personality and vocational problems. Shagass (20) reports the use of word association tests in Canadian pilot selection. The Rorschach has been used for job screening (3), investigation of personality structures associated with particular occupations (12, 18, 19), and selection of mechanical workers (17). Tomkins (22) devotes a chapter of his manual for the TAT to personality diagnosis with respect to work and vocation setting. Projective tests have not been used extensively in vocational guidance for a number of reasons, among them being the training and experience necessary for interpretation, the "excessive" time requirement, and the usual difficulty in relating personality structure to vocational problems in a really useful way. The importance of this latter limiting factor is well illustrated in a study by Kurtz (15) who found that even a specially constructed scoring system did not permit a satisfactory prediction on the basis of Rorschach scores of the success or failure of sales managers.

A more direct approach to vocational guidance problems has been made with interest tests. Berkshire, Bugental, and Cassens (6) report the Kuder Preference Record, the Strong vocational interest blanks, and the California Test Bureau's Occupational Interest Inventory to be the most frequently used. All three tests call for rigidly categorized answers, and scores are interpreted either on a strictly empirical basis in terms of patterns for people already engaged in an

occupation (Strong, Kuder) or a rationalistic a priori basis (California). Although tests of this type work fairly well in practice, they suffer from many weaknesses: (a) the client does not indicate reasons for his preferences; (b) interpretation tends to underemphasize the relationships of the obtained interest patterns to the total personality (8, 14); (c) subtleties of feeling cannot be expressed in terms of the categorized prepared answers; (d) little pertinent qualitative information can be gained by observation during testing; and (e) several studies (9, 13, 16, 21) have shown that answers to such tests can be "slanted" almost at will.

Briefly, it can be generalized that present projective tests do not satisfactorily relate personality to vocational problems, and present interest inventories do not relate vocational problems to personality. Berdie (5) in his review of factors related to vocational interests mentions several studies which report statistically significant relationships between paper and pencil personality tests and vocational interest tests. Such information is interesting and provocative. What seems to be needed is a test which combines the specificity of content now found in the interest inventories with the flexibility and depth inherent in projective tests, and gives information as to the client's interests, their origins, their function in the general personality structure, and how they are likely to affect behavior in the future.

Problem

The purpose of this study was to construct a projective test that would measure vocational attitudes and interests, and at the same time give information concerning related psychological forces operating within the individual's personality. To accomplish this, the following steps were undertaken: (a) construction of drawings of vocational situations which would meet requirements of neutrality, vagueness, and disguised purpose, which nevertheless would evoke a significant variety of responses related to vocations; (b) devising of a scoring system to identify and objectify significant aspects of responses; (c) ascertaining of the reliability of the scoring system; and (d) estimating the validity of the obtained test

scores by determining their capacity to distinguish between groups of known composition, and their consistency with personal data and results from other tests.

Procedure

Materials: The Vocational Apperception Test¹ consists of 18 line drawings, $8\frac{1}{2} \times 5\frac{1}{2}$ inches (1). An individual engaged in a specific occupation is shown in each. Ten of these were designed for administration to women. To facilitate identification, the main figures on these 10 plates are women. The remaining 8 plates were structured for men, with corresponding central male figures.

The ten occupations depicted for women are: (a) laboratory technician, (b) dietician, (c) buyer, (d) nurse, (e) teacher, (f) artist, (g) secretary, (h) social worker, (i) mother, and (j) housewife. The plates for the eight men's occupations show: (a) teacher, (b) executive or office worker, (c) doctor, (d) lawyer, (e) engineer, (f) personnel or social worker, (g) salesman, and (h) laboratory technician. The occupations shown were arbitrarily chosen because of ease of representation and representativeness of occupation.

In drawing the plates, certain rules were carefully followed. Scenes dealt as specifically as possible with one particular occupation or type of occupation. Facial expressions and posturings were ambiguous in feelings expressed. Line drawings were used, since they were felt to be less realistic than photographs, less structured, yet could emphasize the desired aspects of an occupational situation quite adequately.

Before the final sets of plates were drawn, five preliminary plates for five occupations were used to test 15 college men and women informally. Results from this preliminary testing and accompanying interviewing were used to work out the final testing procedure. The test results and discussion indicated that occupational representations should be made less ambiguous and content simpler. This lead was followed in drawing the final set of 18 plates.

Subjects: 40 female and 35 male subjects (Ss) were selected for the two experimental groups.

¹ Henceforth to be referred to as the VAT.

Women Ss were sophomores at Colorado Womans' College to whom 1946 edition Strong Vocational Interest Tests for Women (Revised) previously had been administered as a part of a general testing program for all students. Answers were machine scored for all scales and records showing only one "A" scale rating were segregated from those of the entire sophomore class. Inasmuch as a large proportion of Ss had "A" housewife ratings, in addition to other primary patterns, it was necessary in some cases to disregard the housewife scores. However, Ss selected for the housewife criterion group had no "A" ratings except for housewife. Five women were randomly selected from each of 8 occupational categories thus set up: laboratory technician, dietician, buyer, nurse, teacher, artist, secretary, and housewife.

The 35 men who participated were junior, senior, and graduate students at the University of Denver. Five advanced students majoring in and feeling a primary interest in each of the 7 following fields were chosen:² teaching, accounting, premedical training, engineering, social work, salesmanship, and art. All were given the 1938 Strong Vocational Interest Blank for Men (Revised). Answers were machine-scored for all men's scales.

In the selection of the samples there was no direct control of age or previous occupational experience.

Testing: The VAT was individually administered, responses were recorded verbatim, and manner of responding noted. Es and Ss were matched as to sex, male testing male, and female, female. The following instructions were given:

"The purpose of this test is to find out how people go about understanding human behavior. You probably realize that insight into others' behavior helps us get along with them. Ordinarily, when we meet people, we try to 'size them up.' We do it all the time. I am going to show you some cards and your job will be to tell me a story about the people pictured on these cards. On each card, will you tell me how the person came to be in this

situation, how he (she) feels about it, and what the future holds in store for him (her). Often, it will be necessary to use your imagination."

Following completion of the VAT, male Ss were asked what they thought was being measured. At all times Es' comments were guarded so as not to indicate the purpose of the test, to minimize conscious "slanting" of stories by the Ss.

Scoring: It has already been mentioned that scores were available for all Ss for all scales of the Strong vocational interest blanks. The real need was for a method of scoring VAT responses. It was decided to score for general attitude toward a specified occupation (using S's report and picture content to decide subjectively and arbitrarily which occupation), reasons for entering an occupation, conflict areas or areas of concern, and vocational or personal outcomes. The following outlines cover the scoring system as used by Es and other judges. Scoring consisted of categorizing each response under one or more of the sub-headings in each main scoring area.

A. General attitude toward an occupation (indications from verbalization or judged feeling tone): 1. *Like!* 2. *Like.* 3. *Indifferent.* 4. *Dislike.* 5. *Ambivalence.*

B. Reasons for entering occupations: 1. *Interest and enjoyment.* 2. *Ability.* 3. *Status.* 4. *Income.* 5. *Opportunity.* 6. *Security.* 7. *Altruism.* 8. *Contact with people.* 9. *Seclusion.* 10. *Forced into.* 11. *Excitement or curiosity.* 12. *Experiment.* 13. *Independence.* 14. *Transfer of training.* 15. *Temporary employment.* 16. *Training.* 17. *Contact with field.* 18. *Desire to influence others.* 19. *Idealism.* 20. *Desirable conditions of employment.* 21. *None stated.*

C. Areas of conflict or concern: 1. *Personal conflict (generalized):* a. achievement; b. affiliation; c. aggression; d. inadequacy or insecurity; e. independence; f. recognition; g. valuerative; 2. *Home and parental conflict.* 3. *Marital conflict.* 4. *Financial conflict.* 5. *Educational conflict.* 6. *Vocational conflict.* 7. *Health conflict.* 8. *No conflict mentioned.*

D. Vocational and personal outcomes: 1. *Success.* 2. *Continues in field:* a. mediocre success; b. transfer to better job within field; c. success of no importance; d. no mention of success; e. dissatisfaction in field; f. secures additional

² These fields were chosen because students in them were available. They are those represented on the plates, except that law students and laboratory technicians were omitted, and art students added to the total group.

training. 3. *Leaves field*. 4. *Not clearly stated*. 5. *Confusion*. 6. *Disaster*. 7. *Continuous dissatisfaction*.

For Women Only: 8. *Marry*, but continue in field. 9. *Marry* individual who has an allied interest in field. 10. *Marry* and leave field.

A more detailed account of scoring procedures and criteria can be found in the manual for the VAT (1).

Results

Reliability of scores: When *Es* first attempted scoring with only the outline of the present scoring system, the percentage of rescoring self-agreement ranged from 66 to 75, calculated in terms of the number of times the stories were categorized in the same way in the various scoring areas. This was not felt to be satisfactory, so more precise criteria for differentiation were worked out. After a great deal of discussion of scoring problems coupled with repeated scorings of the same records, standards were set up (1).

Ten protocols were then chosen at random from the as yet undiscussed records and each of two *Es* scored five independently on two occasions a week apart. Rescoring self-agreement was now 86 per cent for all protocols and all scoring categories taken together. Agreement varied little for the various scoring aspects (general attitude, areas of conflict, outcomes, reasons for entering) and there was practically no difference in self-agreement with male and female records. Inspection of the records showed no significant difference between these experienced scorers in reliabilities.

Since it was clearly possible for experienced judges (2*Es*) to attain a high level of personal consistency in scoring, a check on inter-person consistency was made. Four graduate students with experience in TAT scoring³ were given two randomly selected protocols each and a set of scoring instructions. General scoring methodology was explained, and any questions about the scoring categories as given in the instructions were answered. The resulting scorings were compared with previous scorings of the same records by the two experienced *Es*. Although the inter-person con-

sistency between experienced and inexperienced scorers was not as high (mean of 69 per cent) as the personal consistency of the more experienced *Es*, it was felt to be satisfactorily high. There was a wide difference in the scoring agreement of these various outside judges, one showing quite poor agreement (54 per cent) and another agreeing with the experienced *E* essentially as closely (85 per cent) as he did with himself. Scoring agreement was lower for areas of conflict (57 per cent agreement) than for either general attitude (74 per cent) or outcomes (76 per cent). However, scoring aspect had less effect on agreement than the personal scoring proficiencies of different judges.⁴

With several hours training a high level of VAT scoring agreement could certainly be reached. Tomkins (22) points out that "... in TAT workshops it is common for the reliability of ratings to be very low at the beginning but to increase to respectable magnitude with practice."

Comparison of Strong Ratings with VAT Preferences

Each story was scored only once, and only for the occupation the plate was designed to depict. The story was not scored in the few instances (less than 1 per cent) where the occupation was completely misinterpreted. Combining categories, a chi-square test of the independence of Strong scores and VAT preference ratings was made for all *Ss*. The hypothesis of no relationship could be rejected at the 10 per cent level of confidence for the women and at the 2 per cent level of confidence for the men. Although a relationship was thus demonstrated, it is not as clear-cut as one might like. This may be due to the fact that on the whole, although a person may not have many interests in common with people already engaged in and successful in an occupational area, he may be well disposed toward the area as he understands it. Thus it might be ex-

³ Thanks are due Miss Janet Ambler, Mrs. Helen Ammons, Mr. Seymour Levine, and Mrs. Ann Neel for their time willingly spent serving as judges.

⁴ To save space and cost, Tables 1, 2, 3, 4, and 5 have been deposited with the American Documentation Institute. Order Document No. 2748 from the American Documentation Institute, 1719 N Street, N.W., Washington 6, D. C., remitting \$.50 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$.50 for photocopies (6 × 8 inches) readable without optical aid.

pected that the relationship between Strong scores and VAT ratings would not be high.

Characteristics of Responses to the VAT

Reasons for entering the occupation were found difficult to score, so scoring was not done in this area. The first analysis made was of total number of conflicts by conflict area and occupation⁵ for the women and men *Ss* respectively. The men showed more conflicts than women (a mean of 2.0 per story as compared with 1.7). The men were more concerned with achievement, insecurity, vocational, and personal value conflicts; while the women showed more marital and affiliation conflicts. The men showed little overall difference in type of conflict from vocation to vocation, except perhaps for teaching; while women gave evidence of frequent conflicts associated with teaching, but very few in the housewife and mother areas. Although the data are not conclusive, there seems to be some evidence for an interaction between nature of conflict and specific occupation. The women showed more conflicts of aggression and insecurity in the teacher area, and more conflicts of achievement, recognition, and vocational choice in the artist area.

An analysis was made of the outcomes associated with the various plates by the total male and female groups.⁶ The most frequent outcomes in women's stories were success, continuing in the field without success being mentioned, and marrying and leaving the field. Laboratory technicians, buyers, and nurses were pictured as most successful, while dieticians and teachers were described as leaving or desiring to leave the occupation. Nurses, artists, and social workers married, but stayed in the field, and laboratory technicians married some one with an allied interest. Finally, secretaries, dieticians, and nurses were often pictured as marrying and leaving the field.

The most frequent outcomes for men were partial or complete success; it would seem that they must achieve some kind of success. Lawyers, laboratory technicians, and doctors were the most successful; the salesman moved to a better position within the same employment

area; and the teacher left that field of employment.

Ss of both sexes infrequently described teachers as successful, and even where they were successful the success tended to be only mediocre. Both men and women put great emphasis on success, and in addition women frequently told stories about retiring from economic competition to become housewives and mothers.

The final analysis was of the average numbers of words and the associated standard deviations in stories told by the male and female *Ss* to cards picturing similar occupations. Mean lengths were essentially the same (range from 104 to 140 words) for the six pairs of cards,⁷ differing only in the case of teaching where men told significantly shorter stories (mean of 90 words). The story lengths were about the same for all occupational situations except for teacher for men and lab technician for both men and women, which were shorter.

Insight Regarding Purpose of VAT

After each male subject had been tested with the VAT, he was asked what he thought the purpose of the test might be. It was concluded that only 7 of the 35 men came near to understanding the purpose of the test. From the responses given, it could be seen that the likelihood of undesirable systematic slanting to create an impression was small.

Qualitative Evaluation

Much useful information can be gained clinically from qualitative observations and use of "total impressions." Certain observations worthy of mention were made during the VAT testing program. From content analysis of the protocols, it was possible to obtain a fairly definite idea of such things as the extent and accuracy of *S's* information about an occupation, the use to which he would put physical equipment and his feelings toward it, some of his acute personal problems not directly associated with occupations, the place of vocations in his personality dynamics, and how he

⁷ Table 5 is available in the ADI set (see footnote 5). Comparisons are made of lab technician with lab technician, salesman with buyer, doctor with nurse, teacher with teacher, engineer with artist, office worker with secretary, personnel worker with social worker.

⁵ Tables 1, 2, and 3 are included among those available from the ADI (see footnote 4).

⁶ See Table 4 included in the ADI set (footnote 4).

would like to handle his personal problems concerning vocations. It was usually easy to estimate the level of identification with the principal figure in the story, and the degree of personal involvement in the story. Significant personal data of many kinds were obtained, particularly concerning traumatic experiences.

As a rule, when presented with plates concerning their own occupational interests, Ss would show a marked increase in enthusiasm. Ten Ss well known to Es gave clearly recognizable stories. Blind matching would probably have been perfect or nearly so with actual vocational sketches supplemented by personality information.

The following story illustrates the power of the technique: Male, college senior, age 24, story to Plate 4 for men: He's an attorney before a court. On one side you can see the judge; on the other side, the jury. He's defending a guilty person. Like any lawyer, he's looking for any angles to prove the situation. He's not sure of himself, if the circle on the top of his head is a question mark. Am I right? What do you see it as? Why don't you ever tell me if I'm right or not? He's been pushed into this job and can't get away from it. It looks like a pretty tough job. He's not succeeding. He wonders if he's going to change his job or not. He has in mind this problem, and is questioning himself as to what kind of job he will have. He wonders if he is going to succeed or not. It seems to me that he's not sure of himself. He never has been. There is something in back of his mind that he doesn't want to tell people about this. He would rather hide it than tell other people. A kind of non-professional job. He would rather not have too much responsibility.

Discussion

The basic assumption of projective testing is that *S* will interpret stimuli in a way which will reflect the cognitive and emotional organization of his personality. There is a considerable amount of evidence that the VAT provides a suitable situation for the projection of feelings and ideas related to *S*'s vocational problems. Scored conflict areas were different for stories about different occupations, and differed for different individuals. Outcomes

varied with the occupational situation about which a story was told and the sex of *S*. In view of the high scoring reliability these findings seem to indicate a basic validity for the procedure, and are supported in this by a qualitatively observed close correspondence between information derived from personal observation and that from test responses concerning personality facets and vocational problems. The wide variability in the characteristics of the responses given speaks against the hypothesis that the content and structure of the pictures primarily determined the stories told.

The only findings which might be interpreted as indicating a low validity were disagreements in interpretation and a low relationship between Strong interest ratings and VAT attitude-toward-occupation ratings. There were a considerable number of disagreements in interpretation of responses within the scoring system as set up, as evidenced by the failure to obtain more nearly perfect scoring consistency. Analysis of these disagreements almost always led to the conclusion that both interpretations were reasonable, and that they merely represented essentially equally valid but different levels of abstraction in interpretation. Thus the disagreement may be a reflection on rough methods of scoring and inadequate personality theory rather than the validity of the test.

The discovery of anything but a low correlation between Strong ratings and VAT scores would be little short of amazing. Among other things, a high correlation would indicate a close relationship between *interests* common to persons working in an occupational area and the *attitudes* of a group of relatively inexperienced and uninformed persons toward that type of occupation. This relationship is not likely to obtain. With our group, and perhaps any group, one would expect that there would be a wide variety of occupations eliciting favorable responses in a projective test. On the other hand, the Strong items are deliberately weighted to produce score differences between occupational interest areas. What is really needed is a thorough empirical study of personality-vocational-interest relationships.

The VAT is believed by the authors to be a much more versatile clinical instrument than

standardized paper and pencil tests of vocational interest or personality with provisions only for categorical answers. It can be scored effectively to obtain a large variety of significant information about S, including biographical facts of importance, basic conflicts, needs, press, nature of identifications with others of his own sex, methods of problem solution, and attitudes toward occupations and possible reasons for them. Reading and writing are not called for, so it may be useful in growth studies and with relatively illiterate people.

Summary

A set of 18 plates for the projective testing of personality structure related to vocational problems on the college level was constructed, with 10 plates for women, and 8 for men. Each plate was ambiguously drawn but clearly represented a particular occupational area. Methods were developed for identifying and scoring general attitude toward an occupation, reasons for entering an occupation, general and occupational conflicts, and personal or vocational outcomes.

The Vocational Apperception Test (VAT) was administered to 40 female and 35 male college students with primary interests in certain occupational areas as demonstrated by Strong vocational interest ratings or declared major or both. The stories were scored and the scorings analyzed. It was found that consistency of scoring was approximately 86 per cent for experienced scorers rescoring protocols after a week, and 69 per cent between experienced and inexperienced scorers. Indicated areas of conflict and outcomes varied with the sex of the subject, the occupational areas about which the story was told, and the particular subject tested. A low but statistically significant relationship was found between ratings on the Strong scales and the VAT rated general attitude toward an occupation. Information from a follow-up interview indicated that only 7 of the 35 male subjects guessed the purpose of the test.

In the judgment of the authors, the above findings indicate a satisfactory reliability and validity for the test. Its flexible form, and its emphasis on depth information recommend it

for use in the clinical exploration of personal vocational difficulties and in attacks on a wide variety of significant research problems.

Received August 26, 1949.

References

1. Ammons, R. B., Butler, Margaret N., and Herzig, S. A. *The Vocational Apperception Test, plates and manual*. Louisville, Ky.: R. B. Ammons, 1949.
2. Bailey, H. W., Gilbert, W. M., and Berg, I. A. Counseling and the use of tests in the student personnel bureau at the University of Illinois. *Educ. psychol. Measmt.*, 1946, 6, 37-60.
3. Balinsky, B. The multiple-choice group Rorschach test as a means of screening applicants for jobs. *J. Psychol.*, 1945, 19, 203-208.
4. Bell, J. E. *Projective techniques; a dynamic approach to the study of the personality*. New York: Longmans, Green, 1948.
5. Berdie, R. F. Factors related to vocational interests. *Psychol. Bull.*, 1944, 41, 137-157.
6. Berkshire, J. R., Bugental, J. F. T., Cassens, F. P., and Edgerton, H. A. Test preferences in guidance centers. *Occupations*, 1948, 26, 337-343.
7. Bixler, R. H., and Bixler, Virginia H. Test interpretation in vocational counseling. *Educ. psychol. Measmt.*, 1946, 6, 145-155.
8. Bordin, E. S. A theory of vocational interests as dynamic phenomena. *Educ. psychol. Measmt.*, 1943, 3, 49-66.
9. Cross, O. H. A study of faking on the Kuder Preference Record. *Amer. Psychologist*, 1948, 3, 293.
10. Darley, J. G. *Clinical aspects and interpretation of the Strong Vocational Interest Blank*. New York: Psychological Corporation, 1941.
11. Frank, L. K. Projective methods for the study of personality. *J. Psychol.*, 1939, 8, 389-413.
12. Kaback, Goldie Ruth. Vocational personalities; an application of the Rorschach group method. *Teach. Coll. Contr. Educ.*, 1946, No. 924.
13. Kelly, E. L., Terman, L. M., and Miles, C. C. Ability to influence one's score on a typical pencil-and-paper test of personality. *Character & Pers.*, 1936, 4, 206-215.
14. Kilby, R. W. Some vocational counseling methods. 1948. To be published.
15. Kurtz, A. K. A research test of the Rorschach test. *Personnel Psychol.*, 1948, 1, 41-51.
16. Longstaff, H. P. Fakability of the Strong Interest Blank and the Kuder Preference Record. *J. appl. Psychol.*, 1948, 32, 360-369.
17. Piotrowski, Z., Candee, B., Balinsky, B., Holtzberg, S., and Von Arnold, B. Rorschach signs in the selection of outstanding male mechanical workers. *J. Psychol.*, 1944, 18, 131-150.
18. Prados, M. Rorschach studies on artists—painters.

- I. Quantitative analysis. *Rorschach Res. Exch.*, 1944, 8, 178-183.
19. Roe, Anne. A Rorschach study of a group of scientists and technicians. *J. consult. Psychol.*, 1946, 10, 317-327.
20. Shagass, C. Word association tests for pilot selection. *Bull. Canad. Psychol. Ass.*, 1945, 5, 81-82.
21. Steinmetz, H. C. Measuring ability to fake occupational interest. *J. appl. Psychol.*, 1932, 16, 123-130.
22. Tomkins, S. S. *The Thematic Apperception Test; the theory and technique of interpretation*. New York: Grune & Stratton, 1947.
23. Trabue, M. R. The role of the psychologist in vocational guidance. *J. clin. Psychol.*, 1945, 1, 182-185.

Preferred Rate and Extent of the Frequency Vibrato *

John F. Corso and Don Lewis

State University of Iowa

Information has been available for several years on the rates and extents of the frequency vibrato¹ in artistic vocal and instrumental performance. Extensive studies were made by Seashore and his associates on the physical characteristics of vibrato tones. Further, various individuals have expressed opinions concerning the desired rate and extent of the frequency vibrato.

Seashore (6), dealing with the expression of emotion in violin music, reports that "the most beautiful effect is obtained when the pitch oscillation does not exceed one-fourth of a tone." In 1929, Stanley (10) stated that "A proper vibrato must be absolutely regular, must have the correct frequency—about six a second. . . ." Seashore and Metfessel (7) maintain that in vocal performance "An amplitude of approximately a half tone interval in pitch is a good vibrato." In a study on the control and refinement of the vocal vibrato, Wagner (14) selected the rates and extents of recognized artists as the criteria for pleasing vibratos. These opinions, however, were not supported by experimental evidence, except the evidence that certain artistic vibratos had rates and extents of given magnitudes as determined by the physical analysis of the stimulus tones.

Up to the present time, there has been a general lack of information on audience preferences for vibrato tones. Such information should have considerable practical value, both to musical performers and to designers of electronic musical instruments.

Purpose of the Investigation

The present study was undertaken to discover listener preferences for vibrato tones.

* This investigation, which was concerned with the generation and psychological analysis of synthetic music, was part of a research program financed by a grant from the Research Corporation.

¹ The vibrato may be defined as a musical embellishment consisting of a rapid rise and fall in the frequency of a tone (11). This relatively periodic frequency modulation is usually accompanied by synchronous pulsations in intensity and periodic alterations in wave form.

Specifically, the purposes were, first, to determine for musically untrained individuals the preferred combination of rate² and extent³ in the frequency vibrato of a complex tone at each of five octave levels in the equal-tempered musical scale; and, second, to discover the manner in which vibrato preferences varied with the subjects' musical training or native musical ability.

Procedure

Apparatus. As it was highly desirable to employ a musically acceptable auditory stimulus, a cascade of multivibrators (12) was employed as the tone generating unit. Two specific features of this electronic unit made it particularly adaptable to the experimental situation: (a) the multivibrators produce harmonically rich wave forms, commonly exceeding the three-hundredth harmonic in the output (13), and (b) the multivibrators possess the property of easily synchronizing with another voltage having n times its frequency, where n is any whole number. Through the locking action of the set of multivibrators, it was possible to produce a descending series of five tones having a frequency ratio of 2:1. The frequencies of these tones, separated by octave intervals, were 92.5, 185, 370, 740, and 1480 cycles per second. These values correspond to the octave level designations of $F\sharp_2$, $F\sharp_3$, $F\sharp_4$, $F\sharp_5$, and $F\sharp_6$, respectively, at which the preference measurements were made. Inasmuch as the complex tones from the multivibrators were excessively rich in harmonics, it was found necessary to eliminate a considerable number of upper harmonics by means of filters.

The primary component of the experimental apparatus was a frequency vibrato control unit. In this unit, the vibrato rates were

² The rate of a vibrato is the number of pulsations or the number of frequency modulations per second.

³ The extent of a vibrato, as used in this paper, is the width or total range of the frequency modulation and is limited in meaning to the physical measure of frequency in the acoustic wave. Vibrato extent is conventionally expressed as a fraction of a whole step in the equal-tempered musical scale.

controlled by the low frequencies produced by a pentode phase shift oscillator which imparted sine wave oscillation to the multivibrator tone. The rates of oscillation employed were 5.5, 6.0, 6.5, and 7.0 pulsations per second.⁴ The extent of the frequency modulation of the multivibrator tones was controlled by a reactance tube circuit and voltage dividing network connected in parallel with the tuned circuit of the master oscillator. The extents of modulation employed were 0, 0.10, 0.25, and 0.40 of a musical step.

A system of lever action switches, operated by the experimenter, made it possible to combine each of the four vibrato rates with each of the four extents. In this manner, by combining each rate with each extent, twelve vibrato tones and one "straight" tone were produced.⁵ These tones were then presented by the method of paired comparisons, each tone being paired with every other. Regular and inverse orders of presentation were employed.

An electromechanical timing device was arranged to present each of the two stimulus tones of every pair over a high fidelity loudspeaker for a period of 1.7 seconds. The two tones of each pair were separated by an interval of 1.0 seconds. Five seconds elapsed between pairs. At the end of each series of ten judgments, there was a pause of 23.8 seconds. Approximately midway through the experimental session, after 80 pairs had been presented, a five minute rest period was provided during which the subjects were permitted to leave the listening studio.

Subjects. The subjects employed were divided into two main categories: (1) individuals with little or no formal musical training as determined by answers on a personal questionnaire, and (2) individuals with sufficient musical training and ability to be members of the University of Iowa symphony orchestra or chorus. In all, 385 subjects were used. Of

these, 331 were non-trained individuals and 54 were trained musicians.

The non-trained group was further subdivided as follows: (1) Group A, 93 subjects, served at octave level $F\sharp_2$ for the first hour, $F\sharp_5$ the second hour, and took both forms of the Seashore time test at the third hour; (2) Group B, 144 subjects, served at octave level $F\sharp_3$ for the first hour, $F\sharp_6$ the second hour, and took both forms of the Seashore pitch test at the third hour; Group C, 94 subjects, served at octave level $F\sharp_4$ at each of the first two hours and was administered both forms of the Seashore timbre test at the third hour. In all groups, one week elapsed between each of the three experimental sessions.

The trained musicians served only for a single hour at octave level $F\sharp_4$ (370 cycles per second). None of the tests from the Seashore battery was administered to this group.

Prior to the beginning of each laboratory period in which vibrato preference judgments were to be made, the following instructions were read:

"This is an experiment dealing with the properties of the musical vibrato. You will be presented with a series of pairs of tones, each member of every pair possessing a different vibrato or no vibrato. You are asked to select the member of each pair which you like best. If you like the first tone of a pair better than the second, write the number 1 on your record sheet; if you like the second tone of a pair better than the first, write the number 2 on your record sheet. Be sure to make a choice on every pair; that is, write down for each pair either the number 1 or 2. Are there any questions?"

The musically untrained group consisted of experimentally naive persons, all students in an elementary course in psychology. The trained musicians were advanced undergraduate and graduate students majoring in the area of music. Approximately 25 subjects were used at each experimental session. In an attempt to minimize any possible effects resulting from the order of presentation of the pairs of stimulus tones, four different random orders were used at each of the five octave levels tested.

Results

Although the method of paired comparisons was used to obtain the desired data, the mathe-

⁴ The rates and extents of vibrato that were used were selected on the basis of results obtained in a preliminary experiment (I), where it was found that the preferred rates fell between five and seven pulsations per second and the preferred extents between zero and 0.40 of a whole step.

⁵ Pairing four rates with four extents in this case did not produce 16 different tones because zero extent combined with any rate resulted in a "straight" tone (that is, a tone without vibrato).

mathematical analysis did not involve the computation of scale values in accordance with traditional solutions. Instead, rank order scales of vibrato preferences were developed on the basis of the total frequency of preferred judgments for each vibrato tone. Inasmuch as the main purpose of the study was to develop a series of rank order scales from which the preferred vibrato rates and extents could be determined, it was felt that further refinement through a paired comparisons solution was not needed.

At each of the five octave levels tested, a rank order scale of vibrato preference was constructed for the non-trained group on the basis of the total number of preferred judgments made to each vibrato combination at that level. The number of preferred judgments

was then plotted against the vibrato combination of rate and extent to obtain a series of vibrato preference curves.

Figure 1 is a consolidated graph which shows the effect of increasing vibrato extent on the frequency of preferred judgments at each of the five octave levels, when rate is held constant. Since the number of subjects was not equal for all groups, the curves for $F\#_2$, $F\#_4$, and $F\#_5$ were displaced upward by multiplying the frequencies by 1.5. This was to facilitate a direct comparison.

Three features of the graph in Figure 1 are of greatest interest:

(1) The curves are very consistent for all scale positions, with only two slight inversions occurring at a rate of 6.5 pulsations per second. This consistency was further indicated by

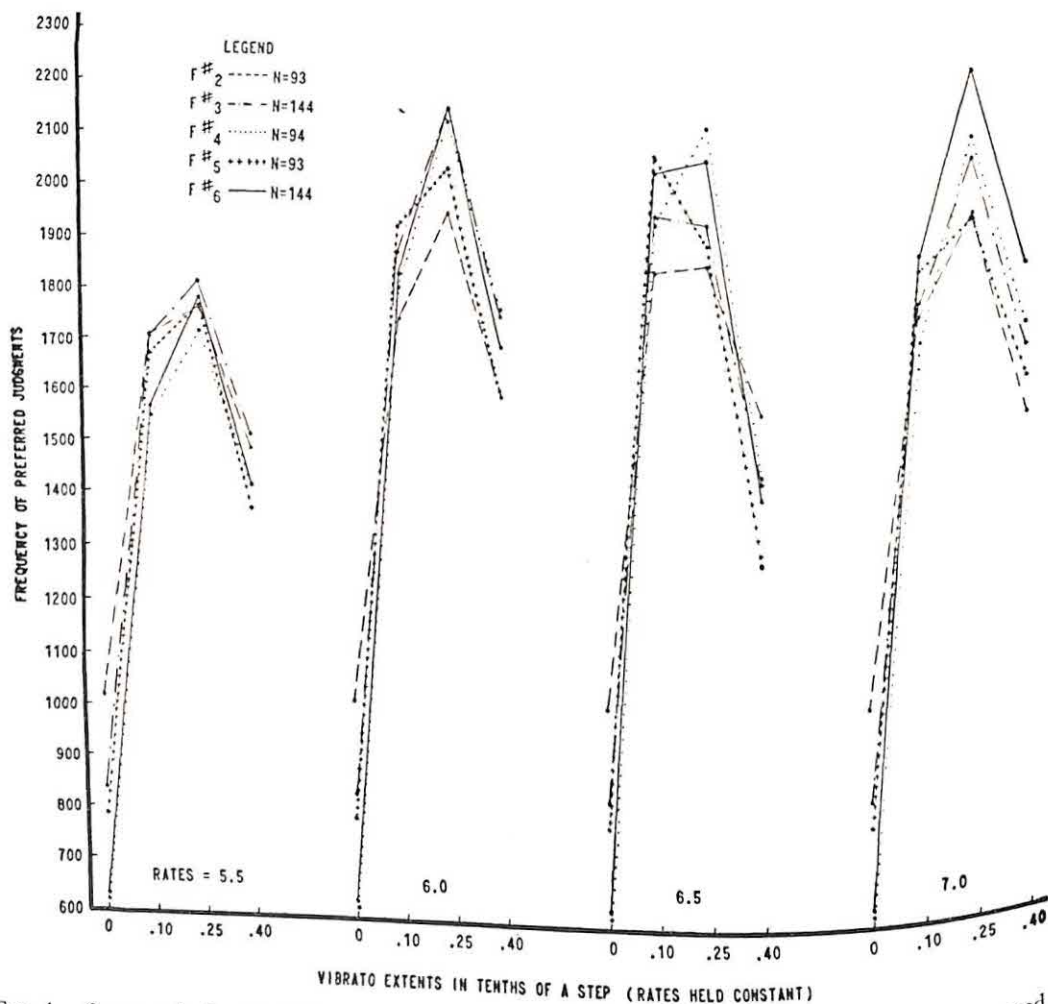


FIG. 1. Curves of vibrato preferences for non-musicians at five octave levels of the equal-tempered scale. (Curves for $F\#_2$, $F\#_4$, and $F\#_5$ were displaced upward by multiplying the obtained frequencies by 1.5. This made adjustment for differences in number of judges.)

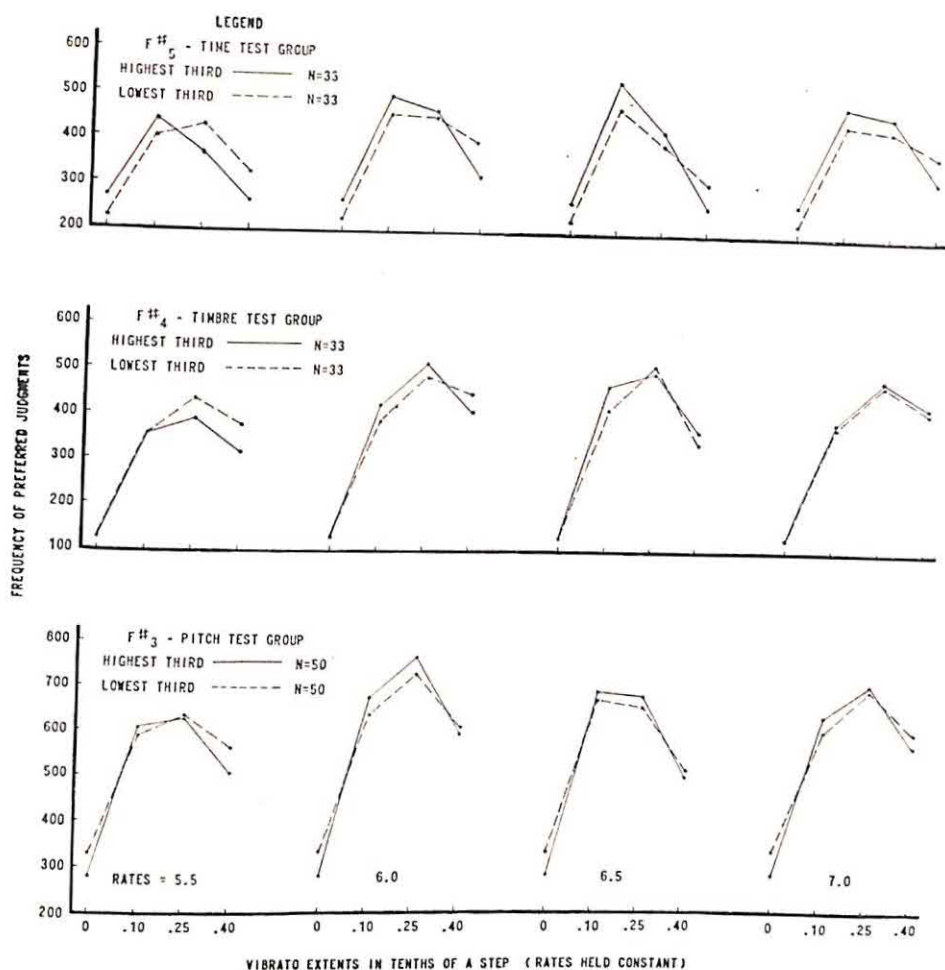


FIG. 2. Curves of vibrato preferences at three octave levels of the equal-tempered scale for groups differing in scores obtained on the Seashore tests.

rank order correlations computed between the scale values obtained at the two different octave levels for each group of untrained subjects. The correlation coefficients were .90 for Group A ($F\#_2$ and $F\#_5$), .91 for Group B ($F\#_3$ and $F\#_6$), and .93 for Group C ($F\#_4$ and retested at $F\#_4$). Since these coefficients were found to be statistically significant, they tended to support the impression obtained from the graphical analysis that vibrato preferences remained constant over a wide portion of the musical range.

(2) With the two exceptions mentioned above, all of the curves indicate a maximum preference for an extent of 0.25 of a step, regardless of the octave tested. The two inversions occurred at octave levels $F\#_3$ and $F\#_5$ where an extent of 0.10 appears to have been slightly preferred when the rate was 6.5 pulsations per second.

(3) The curves for rates of 6.0, 6.5, and 7.0 pulsations per second all reach approximately the same height for a given octave, indicating that these rates were all equally preferred and that extent was the primary determiner of vibrato preference.

In addition to the preference scales obtained at each level for the untrained group as a whole, a series of rank order scales was constructed for several subgroups. These subgroups were obtained by dividing each of the three larger groups approximately into thirds on the basis of scores made on the Seashore tests. Preference scales were then secured for the highest and lowest thirds of each group. Curves derived from the groups taking the pitch, time, and timbre tests are presented in Figure 2. It is apparent from the similarity of these curves that the ability to discriminate time, pitch, and

timbre, as measured by the Seashore tests, was of little consequence in vibrato preference.

A rank order scale was developed for the trained musicians at octave level $F\sharp_4$. The resulting curves, together with those for the untrained group at the same octave level, are presented in Figure 3. Here the curves for the trained musicians were displaced upward by multiplying the frequencies of preferred judgments by 1.7. This took account of differences in number of judges, and facilitates a direct graphical comparison. The curves for musicians show a maximum preference for an extent of 0.10 of a musical step, with the rates of 6.0 and 6.5 pulsations per second equally preferred, while the non-musicians prefer the same rates but a wider extent (0.25 of a musical step).

The chi-square test was applied to determine whether or not the differences between the vibrato preferences of trained and untrained individuals shown in Figure 3 could be attributed to chance factors in sampling. A specific example will be given to illustrate the manner in which these statistical tests were made. The most preferred vibrato tone from

the rank order scale for musicians had a rate of 6.5 pulsations per second and an extent of 0.10 step. The most preferred vibrato tone from the rank order scale for non-musicians had a rate of 6.0 pulsations per second and an extent of 0.25 step. When these two tones were presented to the 54 trained musicians for judgment in a single paired comparison, 36 preferred the tone with a 6.5 rate and 0.10 step extent while 18 preferred the other tone. For the same pair of tones as judged by the 94 non-musicians, only 41 preferred the tone with the 6.5 rate and 0.10 step extent. These frequencies were used in a two by two table and the chi-square value obtained was significant beyond the one percent level of confidence. The hypothesis was rejected that the observed differences were due to fluctuations of random sampling alone. Other chi-square tests for single paired comparisons were made in a similar manner. It was concluded that there was a significant difference between the vibrato preferences of musically trained and untrained individuals. Furthermore, the difference depended upon a different preference in extent of vibrato, not in rate.

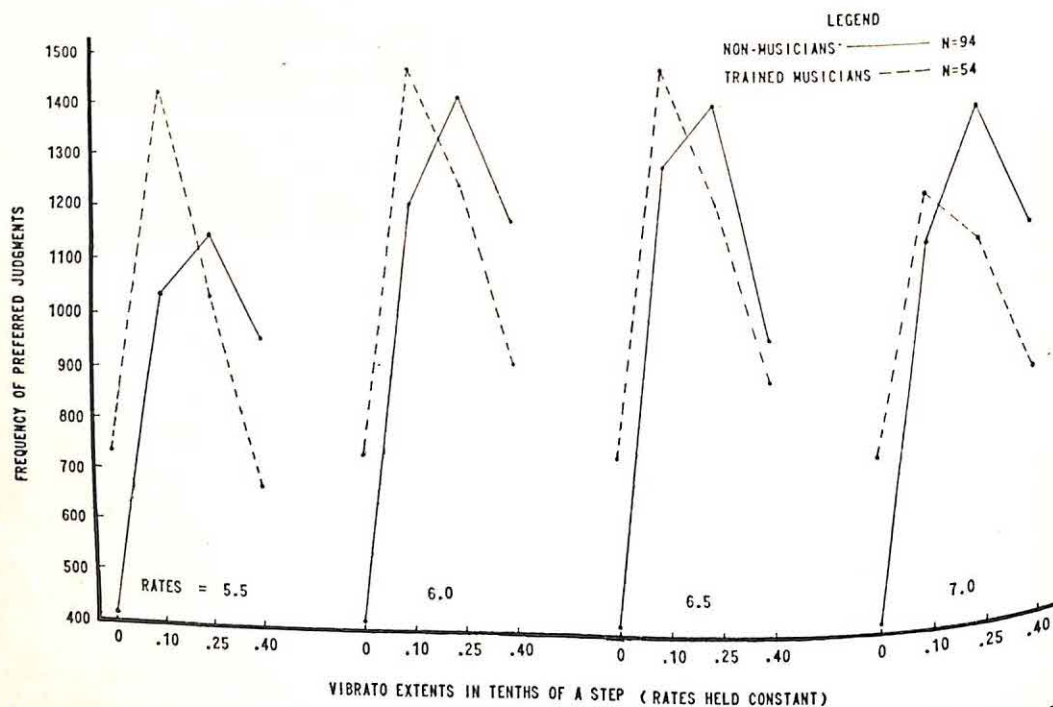


FIG. 3. Curves of vibrato preferences for trained musicians and non-musicians at octave level $F\sharp_4$. (Curves for trained musicians were displaced upward by multiplying the obtained frequencies by 1.7.)

Discussion

The findings reported in this study of vibrato preferences are generally in agreement with the data of previous investigations on the physical characteristics of the vibrato in artistic musical performance. The most preferred rate and extent, as judged by the untrained group of subjects, are similar to the tentative norms established for artistic violin performance by several investigators (2, 5, 9). These authors report the average rate of the violin frequency vibrato to be 6.5 pulsations per second, and the average extent, 0.25 of a tone. Other studies (3, 8) have shown that the typical rates for professional violinists and vocalists are the same, 6.5 pulsations per second, although the violinist's vibrato is only half as wide as the singer's.

Ramsdell (4), employing trained musicians as observers, determined the critical values of rate and extent for maximal richness and for singleness of pitch in a 500 cycle pure tone with a frequency vibrato. The instructions were given, at one time, to increase the rate of modulation until a tone of apparently unitary pitch was achieved, such as would be satisfactory in a single voice. At another time, the subjects were asked to vary the extent of modulation until maximal richness was obtained. The results indicate that the richest, "most unitary" note occurred at a rate of 6.5 pulsations per second and an extent of approximately a semitone. Although the explanation of all the effects of frequency modulation cannot be given at the present time, the experimental evidence seems to support the notion that the close agreement between performed and preferred vibrato rates and extents is dependent upon factors other than those of learning alone.

Summary

1. The purposes of the present study were, first, to determine for musically untrained individuals the preferred combination of rate and extent in the frequency vibrato of a complex tone at each of five octave levels in the equal-tempered scale, and, second, to discover the manner in which vibrato preferences varied with the subjects' musical training or musical ability as represented by scores on the performance tests.

2. The complex tone employed at each of five octave levels was generated by a multi-vibrator unit. This tone was then electronically modulated at rates of 5.5, 6.0, 6.5, and 7.0 pulsations per second. The extents of the frequency modulation were 0, 0.10, 0.25, and 0.40 of a whole musical step. The vibrato tones were presented to groups of subjects for judgment by the method of paired comparisons.

3. Rank order scales of vibrato preference were obtained for two main groups: (1) untrained individuals, and (2) trained musicians. At each octave level tested, preference scales were also obtained for the untrained subjects on the basis of their ability to discriminate pitch, time, and timbre as determined by tests from the Seashore battery.

4. A comparison of the preference scales for untrained individuals indicated that (a) an extent of 0.25 step was preferred over a wide portion of the equal-tempered scale, (b) rates of 6.0, 6.5, and 7.0 pulsations per second were about equally preferred over the same range, (c) native auditory ability had little effect on vibrato preference, and (d) scale values for the retest group showed a high reliability of vibrato preference judgments.

5. The trained musicians tended to prefer the same rates of 6.0 and 6.5 pulsations per second as did non-musicians, but the musicians favored a narrower extent (0.10 of a step).

Received August 30, 1949.

References

1. Corso, J. F. Preferred rate and extent of the frequency vibrato. Unpublished Master's Thesis, State Univ. of Iowa, 1948.
2. Hollinshead, M. T. A study of the vibrato in artistic violin playing. *Univ. Ia. Stud. Psychol. Music*, 1932, 1, 281-288.
3. Metfessel, M. The vibrato in artistic voices. *Univ. Ia. Stud. Psychol.*, 1932, 1, 14-117.
4. Ramsdell, D. H. The psychophysics of frequency modulation. Unpublished thesis, Harvard Univ., 1935.
5. Reger, S. N. The string instrument vibrato. *Univ. Ia. Stud. Psychol. Music*, 1932, 1, 305-340.
6. Seashore, C. E. Phonophotography in the measurement of the expression of emotion in music and speech. *Sci. Mon.*, 1927, 24, 463-471.
7. Seashore, C. E., and Metfessel, M. Deviation from the regular as an art principle. *Proc. Nat. Acad. Sci.*, 1925, 2, No. 9, 538-542.

8. Seashore, H. G. An objective analysis of artistic singing. *Univ. Ia. Stud. Psychol. Music*, 1935, 4, 12-157.
9. Small, A. M. An objective analysis of artistic violin performance. *Univ. Ia. Stud. Psychol. Music*, 1936, 4, 172-229.
10. Stanley, D. *The science of voice*. New York: Fischer, 1929.
11. Stevens, S. S., and Davis, H. D. *Hearing: its psychology and physiology*. New York: Wiley, 1938.
12. Terman, F. E. *Radio engineers' handbook*. New York: McGraw-Hill, 1943.
13. Terman, F. E. *Measurements in radio engineering*. New York: McGraw-Hill, 1935.
14. Wagner, A. H. Remedial and artistic development of the vibrato. *Univ. Ia. Stud. Psychol. Music*, 1932, 1, 166-212.

Book Reviews

Chapanis, A., Garner, W. R., and Morgan, C. T. *Applied experimental psychology. Human factors in engineering design*. New York: John Wiley and Sons, Inc. 1949. Pp. xi+434. \$4.50.

Serious students of man and his work environment will find this book not only provocative but a major contribution in terms of method and technique for the study of conjoint problems. Management personnel, production engineers, industrial consultants, and psychologists will find clear concise statements regarding the use and application of psychological principles to practical problems. The techniques for the solution of many problems which have confounded these groups are clearly set forth.

The book is far from perfect, as the authors would be the first to state. Most of the data and examples are taken from research sponsored by the military services with only isolated instances from industrial situations. This reflects the failure of industry to recognize the importance of such research for their own problems of production and operation. The reader is struck with the wide gaps in the body of human engineering knowledge now available. This book, however, will in all probability stimulate extensive psychological and engineering studies to fill these gaps at the earliest possible moment.

The authors have brought to bear on the problem of the interactions of man, equipment, and his job the research findings of experimental psychologists, industrial engineers, physiologists, and anthropologists. They have presented this knowledge under four broad categories.

1. The effective design of visual displays. This includes a presentation of recent findings on the size, shape, and legibility of letters, numbers, and scale graduations. Evidence is presented on how best to arrange and group displays.

2. The effectiveness of auditory communication of information. Techniques for increasing speech intelligibility and the recognition of

tonal signals are presented together with examples of applications.

3. The effective design of operational controls. This section deals with the optimal control sizes, shapes, types of movement, gear ratios and the resistance of controls.

4. The effective arrangement of individual work places and grouping equipment. This section goes beyond the work of time and motion engineers and concerns itself with the interactions between man and man, man and machine, and machine and machine. These interactions are referred to as "links" and quantitative measurements are developed which permit units to be linked into an integrated system in accordance with the psychological and physical requirements of personnel.

One of the most important sections of the book is the short chapter on the use of statistics in the analysis of errors. Essentially the authors have applied the methods of analysis of variance and of errors of measurement which are widely used in the theory of test construction. The power of these tools and their wide application to other problems is immediately apparent to the reader.

The value of this book lies in the examples of how the methodology of experimental psychology and statistics can be applied to the solution of a wide variety of problems involving men, equipment and the job they have to perform. In the opinion of the reviewer, the writers have rendered a real service not only to psychologists, engineers, and management but also to the man on the job, by summarizing the present status of knowledge in human engineering and indicating how this knowledge can be applied in practical situations.

Jack W. Dunlap

*Dunlap and Associates, Inc.,
New York, N. Y.*

Pease, Katharine. *Machine computation of elementary statistics*. New York: Chartwell House, Inc., 1949. Pp. 239. \$2.75.

"This manual is for students learning to use computing machines in connection with courses

in elementary statistical methods. It is set up to be self-teaching, so that the student, by following the procedures in sequence, may learn to use the machines with a minimum amount of help from the instructor." With this introduction to the preface of her manual, the author has proceeded to outline, in careful detail, the standard calculating machine procedures to be used in addition, subtraction, multiplication, division, and extraction of square roots, and in obtaining the mean, standard deviation, product-moment correlation coefficient, and percentile and standard scores. A separate set of step-by-step procedures in calculation (including appropriate checks for errors) is provided for each of the commonly used models of Friden, Marchant, and Monroe calculating machines.

Practice is also given in complementary numbers, accumulative and subtractive multiplication, multiplication by a constant, and reciprocals, and in the use of tables of products, reciprocals, quotients, squares, and square roots. Standard forms for computing sheets, and a list of 26 references round out the manual's offerings.

The result is an understandable and highly usable manual which should facilitate the learning of calculating machine procedures either by the student formally enrolled in an elementary statistics course, or by the individual who wishes to learn these techniques on his own. Since it limits itself to the mechanical aspects of computation, it complements, rather than supplants, the more conventional elementary statistics text.

Kenneth E. Clark

University of Minnesota

Boynton, Paul W. *Selecting the new employee*. New York: Harper and Brothers, 1949. Pp. 136. \$2.00.

This is a practical, down-to-earth review of the major principles of employment selection written in a simple, readable, non-technical manner. It is particularly well suited to use by executives interested in establishing an employment department or by beginners in employment work who are seeking orientation in this field.

The author places justifiable stress on the

obligation of the employer to his employees and emphasizes the importance of proper selection and placement. He next discusses the qualifications of the employment man. This is followed by an outline of sources and methods of recruitment, with particular reference to college recruiting. Next follows a discussion of the functioning of the employment department, which is probably one of the best parts of the book. In a manner which reflects the author's long years of experience in practical employment work, he brings out the importance of building acceptance for the employment department and indicates many ways in which this can be done. The balance of the book is devoted to an exposition of interviewing techniques, a brief discussion of tests and a short treatment of induction and training.

Probably one of the most valuable contributions of the book is the author's healthy and realistic appraisal of the contribution of psychological tests to employment work.

If one were disposed to criticize the book, it would be chiefly on two counts: the first is that it contributes little that is new or unique. It is simply a popularly and clearly written handbook of sound (but far from all-inclusive) employment procedures. It is obviously written by a man who has had a wealth of practical employment experience but is not too well acquainted with the literature nor with the more sophisticated developments in the field.

Robert N. McMurry

*Robert N. McMurry and Company,
Chicago, Illinois*

Bennett, George K., and Cruikshank, Ruth M. *A summary of clerical tests*. New York: The Psychological Corporation, 1949. Pp. iii+122. Paper, \$1.25.

This booklet is "an attempt to bring together in a single publication pertinent information regarding tests used in selecting and upgrading clerical workers"; it is a companion to an earlier booklet dealing with manual and mechanical ability tests. There are two distinct parts: a discussion of the development and use of tests of clerical ability, and a series of brief descriptions of specific tests.

An historical sketch of the growth of clerical occupations and survey of developments in

clerical testing from 1912 until World War II are followed by a review of the types of items used in clerical tests, material which could be very useful to persons seeking ideas for clerical test construction.

Two chapters are devoted to job descriptions of clerical occupations and to reviews of studies in the selection of workers for these occupations; these are helpful as an overview of what has been tried, but they impress this reviewer as being insufficiently analytical and integrative. That some attempt was made along these lines is illustrated by the authors' noting of the fact that Hay's Pennsylvania Co. norms for the Minnesota Clerical Test are very similar to those compiled in the USES work of Stead, Shartle, and others. But in general the criteria of success are not carefully examined (e.g., on page 26, Thorndike's use of earnings at ages 20-22, when beginning workers have not yet had a chance to prove their worth and to earn accordingly) and sometimes they are not even described (e.g., Oberheim's "library performance" criterion on page 24): a serious defect in view of the recent emphasis on the need to validate the criterion. Sometimes results which might mystify or even disillusion the unsophisticated are not discussed, as in the case of the perfectly logical *negative* correlations between Otis scores and salary cited in Table V, p. 28.

Chapter VI is a useful discussion of who should be tested, when, and by whom: content often omitted in the more academic treatises, but especially necessary in those read by laymen.

The test summaries which take up 45 of the 115 pages of text give publication and purchasing data excepting cost (wise in these inflationary days), administration and scoring methods and time, type of content, reliability, validity, norms, and some evaluative comments. These descriptions range in length from one-half page to about one and one-half pages for each of 32 tests. The descriptive material is more adequate than the evaluative, for while the authors have given as their objective the inclusion of "a maximum of objective information regarding . . . reliability and validity," this information ranges from "No studies reported," through "Correlation coefficient of .63 between speed scores and super-

visor's ratings," to somewhat more detailed data not exceeding seven lines.

The analytical comments sometimes do something to remedy this paucity of detailed objective evidence in the test summaries, as when it is stated "The manual does not define the superior group as to number of cases or number of steps in the rating scale" (p. 99), or when the authors say (of one of Bennett's clerical tests), "It needs further study to show its usefulness in differential placement work" (p. 88), or "This is another among the recently published tests (in this case not Bennett's) which include some normative data but little except face validity or item types known to be valid to indicate its areas of usefulness" (p. 99). But these comments are brief, and often non-evaluative. For example, concerning one test which undeservedly (according to the authors' own criteria) receives a full page of description, the authors write simply that it is an American revision of an English test, that the author points out that good secretaries do well on the whole test whereas routine office workers do poorly on certain subtests, and that scores increase with schooling and are probably related to intelligence. They do not state that the American validation of this test is extremely limited (only 3 of 5 references have *any* such data) and that the mentioned lack of occupational norms makes it useless in vocational counseling or selection.

Thirty-two generally available tests are described in the manner just discussed; 8 tests, the use of which is restricted, and another 34 tests which are no longer available or virtually unused, are also described more briefly. While it is admittedly difficult to categorize tests in this manner to the satisfaction of all readers, this reviewer is inclined to believe that Bennett and Cruikshank would have rendered a greater service to personnel psychology if they had been more analytical and evaluative in the discussion of each test, or if they had applied more rigorous standards in deciding which tests should be treated. One test publisher has told the reviewer in conversation that the mere mention of a test in a book on tests increases the sale of that test: if this is so, then *a number of tests which have already been tried and found wanting are likely to receive a new lease on life from this publication.* It might

have been a greater service to personnel psychology and vocational guidance if Bennett and Cruikshank had given more space to the tests which the evidence shows to be most worth using, and had relegated to a small-type appendix all of the less promising tests which needed to be discussed for historical reasons or for completeness of coverage.

This raises the question of the readers for whom the authors were writing, of the intended and probable use of the booklet. *Personnel psychologists* will find this summary a time-saving survey of the field, a very valuable source of leads to studies they should be familiar with, and a helpful reminder of what is available in the way of tests; they will also find that the test summaries do not enable them to make final judgments about tests, but that they tell enough to indicate what is worth looking into more intensively both in the literature and in their own research. *Vocational counselors* will also find the survey and the leads helpful, but will need to go to Buros' yearbook and to the more intensive treatises of specific tests found in some texts. *Personnel managers* and other executives, who may have had a little training in measurement but who are not psychologists, are likely to be confused by the large number of briefly described and sketchily evaluated tests. As this last is one of the largest probable consumer groups, and also one which is likely to make least use of the literature to which this booklet could lead them, this limitation becomes even more important. Bennett and Cruikshank are not the first authors who have, because of the very nature of their field, written for too heterogeneous an audience; but it would be interesting to see what they would produce, if they wrote two versions of the booklet, one for technicians and counselors highly trained in testing, and one for personnel workers and counselors with at most a course or two in measurement.

In closing, the reviewer, whose eye tends to be somewhat jaundiced when reading test discussions by test authors and publishers (*timeo Danaos, et dona ferentes*), would like to point out that Bennett and Cruikshank have conscientiously treated their own tests as

objectively as those of other authors and publishers.

Donald E. Super

*Teachers College,
Columbia University*

Weitzman, Ellis, and McNamara, Walter J. *Constructing classroom examinations—A guide for teachers*. Chicago: Science Research Associates, 1949. Pp. xvi+153. \$3.00.

Of all the guide books designed to assist teachers in building effective classroom examinations this is perhaps the most elementary. Beginning with validity and reliability and ending with the statistical analysis of test scores, it covers in simple, non-technical language the customary topics, briefly and superficially. Its greatest value will accrue to those teachers or prospective teachers who know nothing about objective test construction and who desire to learn but little.

Those who read this book for the purpose of becoming informed about important advances in achievement test construction during the past twenty years will be disappointed. Validity and reliability are defined and methods of measuring them given but the reasons why they are important and the factors which influence them are neglected. Instead of emphasizing the place of course objectives in the process of item construction and sampling, we find recommended a topical outline of subject matter with textbook page references. No instructions are given regarding ways of organizing content and objectives to facilitate item construction. Many of the test items used as examples are excellent but there is a preponderance of the more factual type. The teacher-made answer sheet and perforated scoring key is not treated, but the consumable type of test with panel scoring key is. Eight pages are devoted to a cumbersome procedure of item analysis designed to yield a measure of the difficulty of the items. One page is devoted to a simple index of discriminating power without emphasizing its importance or the factors which influence it. Percentile ranks are not computed from the mid-frequencies of the scores. The student learns to compute

means and standard deviations but standard scores are not mentioned. Certainly, better manuals on the construction and use of achievement tests have been available for many years.

Walter W. Cook

University of Minnesota

Cavan, R. S., Burgess, E. W., Havighurst, R. J., and Goldhamer, H. *Personal adjustment in old age*. Chicago: Science Research Associates, 1949. Pp. xiii+204. \$2.95.

The distinctive contribution of this book to applied psychology lies in its detailed account of the development, testing and application of an Attitude Inventory and an Activity Inventory for the study of persons past sixty years of age. The aim of these inventories is to secure data on activities and attitudes in various areas including health, family, friends, work and economic security.

In testing the validity of these inventories, interesting auxiliary schedules were developed: a check-list of personal characteristics which an interviewer might observe, a set of word portraits, and a list of symptoms supposed to indicate senility (Appendixes D, E, F).

More than 8,000 schedules were mailed out. More than half of them went to retired teachers, retired ministers and widows of ministers. More than a quarter of them went to sociology professors who distributed them through their students. Usable schedules returned by mail numbered 2743, and these were supplemented by 245 schedules obtained by interview (pp. 171-172).

Table 30 (p. 134), based on the entire group of 2,988 responses to the Attitude Inventory, shows the correlations of partial scores (indicating degrees of adjustment in different areas) with one another and with total scores. Attitudes toward leisure showed highest correlation with the total score in both men and women (.73 and .70); happiness, feeling useful, and satisfaction in work came next; religious attitudes had the lowest correlations (.35 and .29).

On the position of religion the authors make the following comment: "This is not surprising, for almost a third of the subjects were ministers or their wives, and they probably show relatively little variation in religious attitudes while they show a wide variation in other attitudes" (p. 133).

Tables 7 to 17 (pp. 48 to 59) are based on a smaller "study group" of 499 men and 759 women (pp. 46-48 and Appendix C). The relation between the "study group" and the total group is not made clear in the text. Replying to a query from this reviewer Dr. Havighurst states: "The 'study group' consisted of all the respondents except the two major occupational groups, namely the retired teachers and the retired ministers and their wives. Thus, the 'study group' consists of the people who are described on pages 170 and 171 under paragraphs (d) and (e), and the groups described in the first few paragraphs on page 171."

Trends with age found in the "study group" include (p. 60): Increased feeling of economic security in spite of lowered amount of income. Increase in religious activities and dependence on religion. Decrease in feelings of happiness, usefulness, zest, and a corresponding increase of lack of interest in life.

Sex differences include the following (p. 61): Women feel somewhat more secure economically than men. Women report more physical handicaps, more illness, more nervous and neurotic symptoms, and more accidents; they feel less satisfaction with their health than do men. Women have more religious activities and more favorable attitudes toward religion than do men. Women are less happy than men.

The authors have made an exceptionally valuable contribution by their thorough, cautious, and critical development of the two inventories and by using them in a study of a large number of cases.

Albert R. Chandler

Ohio State University

New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota

- The individual and his religion.* Gordon W. Allport. New York: The Macmillan Co., 1950. Pp. 147. \$2.50.
- The supervisor's management guide.* Louis Baldwin et al. New York: American Management Association, 1950. Pp. 200. \$3.50.
- Know yourself. A workbook for those who stutter.* Revised edition. Bryng Bryngelson, Myfanwy E. Chapman, and Orvetta K. Hensen. Minneapolis: Burgess Publishing Co., 1950. Pp. 159. \$2.00.
- Making work human.* Glen U. Cleeton. Yellow Springs, Ohio: Antioch Press, 1949. Pp. 326. \$3.75.
- Occupational therapy.* William R. Dunton, Jr. and Sidney Licht, Editors. Springfield, Ill.: Charles C Thomas, Publisher, 1950. Pp. 350. \$6.00.
- A handbook of employment interviewing.* John M. Fraser. London: Macdonald and Evans, 1950. Pp. 212. 8/6d.
- Theory and practice of psychological testing.* Frank S. Freeman. New York: Henry Holt and Co., 1950. Pp. 518. \$3.50.
- Fields of psychology.* Second Edition. J. P. Guilford, Editor. New York: D. Van Nostrand Co., Inc., 1950. Pp. 779. \$5.00.
- A handbook of applied psychology.* Douglas H. Fryer and Edwin R. Henry, Editors. New York: Rinehart and Co., 1950. Two Volumes, pp. 826. \$12.50.
- Counseling adolescents.* Shirley A. Hamrin and Blanche B. Paulson. Chicago: Science Research Associates, 1950. \$3.50.
- Learning and instruction.* National Society for the Study of Education, Forty-Ninth Yearbook, Part I. Nelson B. Henry, Editor. Chicago: University of Chicago Press, 1950. Pp. 352. \$2.75.
- The education of exceptional children.* National Society for the Study of Education, Forty-Ninth Yearbook, Part II. Nelson B. Henry, Editor. Chicago: University of Chicago Press, 1950. Pp. 400. \$2.75.
- Situational factors in leadership.* John K. Hemphill. Columbus: Bureau of Educational Research, Ohio State University, 1949. Pp. 135. \$3.00, cloth; \$2.50, paper.
- Child development.* Second edition. Elizabeth B. Hurlock. New York: McGraw-Hill Book Co., Inc., 1950. Pp. 669. \$4.50.
- How to be happy though young.* George Lawton. New York: The Vanguard Press, Inc., 1949. Pp. 300. \$3.00.
- The science of chance.* Horace C. Levinson. New York: Rinehart and Co., Inc., 1950. Pp. 348. \$2.00.
- The meaning of anxiety.* Rollo May. New York: The Ronald Press Co., 1950. Pp. 376. \$4.50.
- The culture of industrial man.* Paul Meadows. Lincoln: University of Nebraska Press, 1950. Pp. 216. \$3.75.
- Job evaluation.* John A. Patton and Reynold S. Smith, Jr. Chicago: Richard D. Irwin, Inc., 1950. Pp. 338. \$4.50.
- The envelope.* James S. Plant. New York: The Commonwealth Fund, 1950. Pp. 299. \$3.00.
- Introduction to psychosomatic medicine.* C. Alberto Seguin. New York: International Universities Press, 1950. Pp. 320. \$5.00.
- How to make achievement tests.* Robert M. W. Travers. New York: The Odyssey Press, 1950. Pp. 180. \$2.25.
- Human relations in modern industry.* R. F. Tredgold. New York: International Universities Press, 1950. Pp. 192. \$2.50.
- The development of a test for selecting research personnel.* Manpower Branch, Human Resources Division, Office of Naval Research. Pittsburgh: American Institute for Research, 1950. Pp. 33.

Journal of Applied Psychology

VOL. 34, No. 4

AUGUST, 1950

The Adequacy of Employee Selection Reports

Margaret Hubbard Jones

University of California at Los Angeles

An examination of well over 2,100 references from the obtainable world literature on employee selection has permitted an analysis of practices in both experimental design and report, the results of which are perhaps surprising. These references cover the period 1906 to 1948, within the ability of the author to locate the references and within the ability of reference librarians to locate copies of unusual and foreign periodicals and monographs. There are further limitations to the data to be analyzed here which were imposed by the primary purpose of the literature search. That aim was the compilation of abstracts of employee selection reports which should contain the actual data presented, together with sufficient information to enable the reader to evaluate the study without referring to the original report. The work was prompted by the difficulty experienced in this particular field in locating the widely scattered references (we had reference to more than 300 separately-titled periodicals—and many volumes of each one—as well as books and monographs) and by the fact that most industrial psychologists do not have the time or facilities necessary to review this literature. These abstracts appear elsewhere (2).

Due to the large volume of material in this field and to the fact that many articles which seem important contain virtually no information, it did not seem economic to abstract all possible references, and the survey is thus limited to those studies which can be evaluated: those in which relatively complete validation data are presented, together with specific tests used, N and job studied. Further, since we were interested in selection of employees for industrial concerns we have also excluded the special fields of selection for the armed forces

and pilot training as posing special problems. It has by now become abundantly obvious that seemingly slight changes in working conditions, incentives and parent population, to mention the more obvious factors, may result in the failure of even a carefully executed selection program. In view of this, it seemed wiser to exclude those studies in which the criterion was school grades or teachers' ratings, and those carried out on military personnel even where the jobs are similar to civilian jobs, because of the differences in motivation and working conditions. After we have eliminated the reports which are, by our definition, special problems and those which are so inadequately presented or executed that they cannot even be evaluated by the reader (a large proportion of the total) there remain 427 reports, or 20% of the total number of references. In this analysis we shall be concerned entirely with these 427 reports which can be evaluated. Since many articles report results on diverse jobs, very often with different tests and different statistical procedures, we have at times referred to the number of separately-treated groups rather than to the number of titles and have endeavored to make the distinction clear wherever the former occurs.

These 427 references are largely American—about 80%—both because of the greater availability of privately published and unpublished material of American origin and because of the larger total volume of articles. It does not appear that the percentage of acceptable articles is much larger for American work than for that of any other country.

The volume of acceptable articles is shown by year in Figure 1. The slow rise to a peak after World War I, followed by a decline through the depression era is not unexpected.

Whether this can be entirely attributed to over-selling of testing, as is usually done, or whether it does not also reflect to some extent the general decline in business activity is a debatable point. The annual volume of articles reached a high again in 1941, as business was recovering, fell off, understandably, during the war years, and has now reached an all-time high. Whether it will remain high probably depends partly on the quality of current work and partly on the general level of industrial activity.

The ten jobs which have been most frequently studied are as follows: Salesmen, 75; Clerical Workers, 60; Teachers, 49; Assemblers, 23; Executives, 23; Inspectors, 23; Supervisors, 21; Typists, 17; Stenographers 14; and Machinists, 9. This order does not necessarily reflect either the importance of the job or the difficulty of selecting good workers. Among the jobs which appear to have been acceptably reported but once are brick-layers, grocers, scientists and deans of women. As can be seen, salesmen lead the list. This, of course, includes salesmen of all sorts and many of the jobs are quite different. The same comment applies to the other categories. There is

usually no way of determining from the published reports whether the jobs in two studies are comparable. As Ghiselli has shown, there is an astounding range of reported validity coefficients for the same general type of test in any broad occupational classification (3). For clerical occupations he found the range to be over .90. There are many reasons for this state of affairs but one which has perhaps not been sufficiently emphasized is the lack of adequate job description in published reports. It is now fairly generally recognized that a selection program is very much situation-bound but the corollary, that this requires precise job description, is not currently practiced.

Let us now examine in more detail the 427 reports which represent the cream of the crop. The number of subjects used in the investigation is an important factor in determining the predictive value of the results. Except in the majority of cases using less than 20 subjects, the N in and of itself does not tell the whole story. Much depends upon how the data are treated and whether or not the total population of employees on a particular job, or a rep-

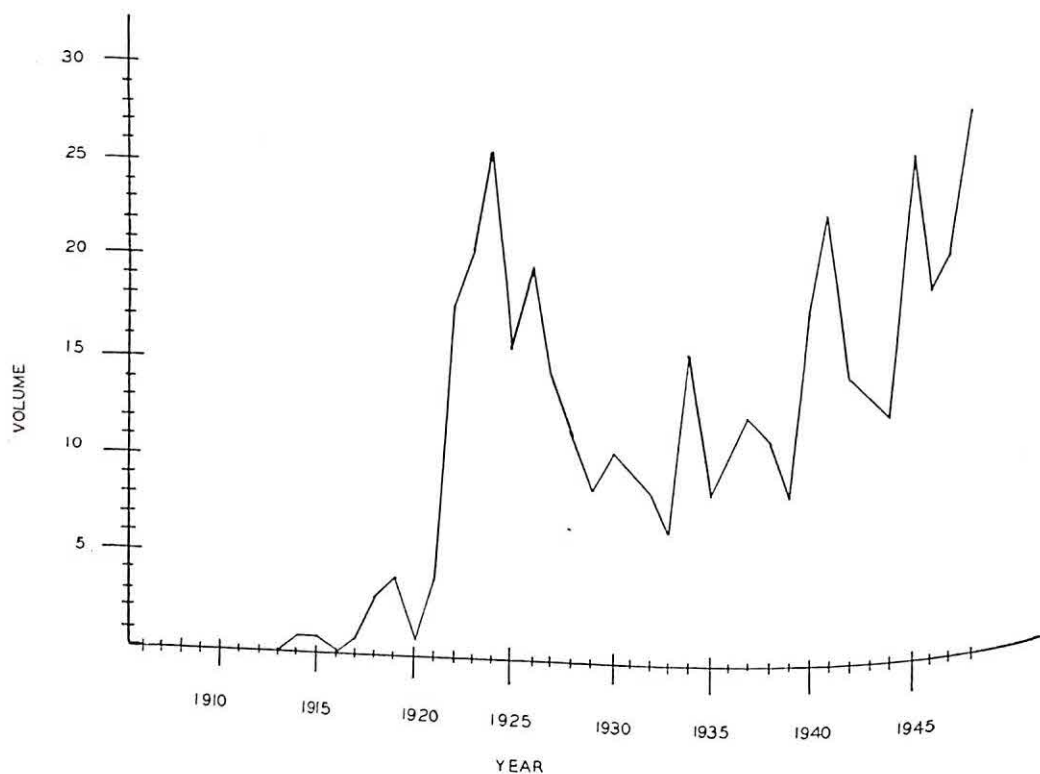


FIG. 1. Number of employee selection reports by year.

representative sample thereof, was used. Nevertheless, it is instructive to analyze the trend in this respect. The results of the analysis by number of subjects is as follows: Less than 10, 17; 10-19, 97; 20-29, 93; 30-49, 129; 50-99, 188; and 100 and above, 257. Even more unexpected than the number of groups with small N is the number with 50 or more and the 257 groups containing 100 or more subjects. The latter are by and large the more recent studies and the trend is encouraging.

Statistical Techniques

An analysis of the statistical techniques used for presentation of the results of validation procedures is interesting, but again the particular statistic used does not guarantee adequacy of treatment because the assumptions governing its use may not have been met and the statistic best suited to a given problem may not have been chosen. Table 1 shows the frequency with which various measures are used. Correlational techniques are the most popular, accounting for 285 out of 525 separately-treated groups. Of these only 172 give measures of significance, and although they can be calculated from the data provided, it is safer, considering the heterogeneous nature of the audience in this field, to present the standard errors along with the coefficients of correlation. Furthermore, the author has the real responsibility for complete presentation of all the statistics necessary to an interpretation of his investigation. Occasionally one even finds an author concluding that the correlations reported have clearly shown a relationship be-

tween test scores and criterion whereas actual calculation shows the correlations to be not significantly different from zero. This practice is general and no single individual should shoulder the blame for it.

Group comparisons of various sorts account for 185 cases, but of these only 28 include measures of the significance of group differences (although sometimes such measures could be calculated by the reader). Group comparisons may take such forms as: differences in mean test scores between the upper and lower 50% of employees as judged by the criterion, or average test scores for groups judged best, average and poorest by their supervisors (many times without N or sigma for each group being indicated), or per cent of those scoring within certain limits who were judged good as against the per cent who were judged poor, etc. Only occasionally are critical ratios or *t*-ratios or similar measures included. The importance of testing results for significance cannot be over-emphasized. In the case of group comparisons it is more serious than where correlation is used because in most cases there are not sufficient data to enable the reader to perform the proper calculations for himself. In one particular case, where the author concluded that his tests were efficacious for selection but neglected to supply any measure of the significance of the differences found between groups, calculation of the significance of percentage differences (the only data available) showed the differences to be exceedingly insignificant.

In 55 cases there is incomplete statistical analysis. In a few cases the raw data are presented with no summary statistics. In many instances we find the results expressed only as "per cent agreement" between test scores and criterion scores, or a brief statement that a critical score of a certain magnitude would have eliminated a given per cent of the poor group and ordinarily a smaller per cent—although we do not know that it is a reliably smaller per cent—of the good group. In 5 cases the authors are content to present graphs alone, sometimes with the differences very much exaggerated by the scale and baseline chosen.

One gains the impression that many times, even where adequate statistics are used, the basic requirements for their use have not been

Table 1

Statistical Measures Used for Validation

Measure of Correlation	Number of Groups So Treated
<i>r</i>	136
Rho	94
R	35
<i>r</i> _{bis}	7
tetrachoric	8
other	5
Total	285
Group Comparison	185
Inadequate Treatment	55

met.¹ One should be able to assume, for example, that when an r is reported the conditions for its proper use have been met, but in view of the inadequacy of many of the statistical treatments one cannot always so assume. A more obvious criticism of many studies is the manner in which subjects are selected. In spite of the fact that the assumptions underlying many of the statistics used require a reasonably random sample, biased rather than random sampling seems to be the rule. It is a common procedure to select certain employees to serve as subjects but rarely are we given any information which indicates that the sample was a selected one or how it was selected. A frequent practice is the artificial creation of a heterogeneous experimental group by the use of only extreme employee groups (the upper and lower 25%, for example), a practice which may spuriously raise the validity coefficients.

A point too often overlooked is that a selection program is intended to select among *applicants*, not among employees, and the two groups are not identical (cf. 5, 7, 10). "Natural selection" on the job—the survival of the fittest—operates to make the employee group more homogeneous than the applicant group. This may spuriously lower validity coefficients and change critical scores. Further, the employee group will often not show a normal distribution in a trait which is highly correlated with ability to produce on the job and since in most industrial situations it will be impossible to correct for this error, the usefulness of the employee group as a basis for a selection program is further limited. The best practical solution to both problems—that of bias in sampling and that of restriction of range in employee groups—seems to be the use of two groups: first, a randomly selected employee group as a trial group, for reasons of economy, and second, an unselected applicant group as a follow-up group, to discover whether or not the selection program will select among appli-

cants as well as among employees. This solution, the use of separate groups, has the further advantage of permitting a pragmatic estimate of the shrinkage in multiple correlation. This is a real advantage. An example is Selover's two samples of clerical workers ($N=193$ and 85 , respectively) which yielded multiple correlations for 4 tests and criterion of .41 and .33, respectively (9). An instructive example of the danger involved in putting one's faith in a single small sample, particularly when that sample has been used to develop a scoring procedure, is given by Kurtz (4). Here a scoring technique for the Rorschach Test was developed which classified correctly 79 out of 80 sales managers. This was so impressive to many people concerned that they were prepared to start using it as a selection device immediately. A follow-up on a second sample yielded a validity coefficient of .02!

The criterion is, of course, a question of utmost importance but we cannot discuss all the ramifications of the problem here. Entirely aside from the question of the applicability of the criterion as a real measure of job success—the validity of the criterion—we find a problem in the reliability of the criterion. Only 95 reports, or 22% of the 427 acceptable reports, make some attempt to include measures of the reliability of the criterion, and yet it has a profound influence upon the results of the validation procedures. Of course, low reliability will not give spuriously high validities—rather the opposite—but many studies appear to lead to the conclusion that certain tests are worthless in a given situation, whereas low criterion reliability may actually be masking a significant relationship. Surely, if a study is worth doing at all, the reliability of the criterion should be ascertained.

Another difficulty in connection with the criterion is the operation of external influences such as age, experience and length of time on the job.² Unless cognizance is taken of these variables the results are difficult to interpret, to say the least, and few studies control any of these factors. For example, it is easy to see

¹ The criticism of the use of statistical methods in research on the Rorschach Test by Cronbach (1) may be applied in part to employee selection research even where other tests are used. Especially to be noted are his criticisms of the selection for emphasis of a few "significant" differences from among many insignificant ones, whether the comparisons are explicitly made or merely implied, and his insistence on the use of a second independent sample so that chance variations in test scores will not be given undue weight.

² An attempt to compensate for bias in the direction of longer service may be found in a recent study by Rundquist and Bittner (8), and McMurtry and Johnson attempted to secure ratings on their subjects after approximately equal time on the job (6).

how age may be predictive of job success if age is influencing the criterion either in its own right or operating through length of service, and yet most studies lump together not only all age groups but employees with widely differing lengths of service. On the other hand, if age or length of service is influencing the criterion a significant relationship with test scores may be masked. One often suspects a further contaminating factor when ratings are used if test scores are not kept strictly confidential until after ratings are made. The facts seem to indicate that more attention must be paid to proper experimental design if the results of selection studies are to be useful.

How many reports, then, meet all criteria of adequacy in both experimental design and report? We find that 46 out of the 427 originally selected contain no second or follow-up group but are acceptable in all other respects, such as sufficiently large N, adequate and complete statistical presentation throughout, etc. We further find that 17 are adequate in all respects except that no measure of the reliability of the criterion is presented. Finally, if we count the total number of reports which are satisfactory in all respects we discover only eight, or .4% of the 2100 references with which we started. These eight studies are as follows:

1. Bellows, R. M. Studies of clerical workers. Chap. VIII in Stead, W. H., Shartle, C. L., et al. *Occupational counseling techniques*. New York: American Book Co., 1940, ix + 273, pp. 144-146. (Study of coding clerks.)
2. Blum, M., and Candee, B. The selection of department store packers and wrappers with the aid of certain psychological tests. *J. appl. Psychol.*, 1941, 25, 76-85.
3. Guilford, J. P., and Comrey, A. L. Prediction of proficiency of administrative personnel from personal history data. *Educ. psychol. Measmt.*, 1948, 8, 281-296.
4. Holliday, F. The relation between psychological test scores and subsequent proficiency of apprentices in the engineering industry. *Occup. Psychol., Lond.*, 1943, 17, 168-185.
5. Otis, J. L., Endler, O. L., and Kolbe, L. E. Data-analysis methods. Chap. VII in Stead, W. H., Shartle, C. L., et al. *Occupational counseling techniques*. New York: American Book Co., 1940, ix + 273, pp. 113-136. (Study of department store salespersons.)
6. Rundquist, E. A., and Bittner, R. H. Using ratings to validate personnel instruments: a study in method. *Personnel Psychol.*, 1948, 1, 163-183.

7. Sartain, A. Q. Relation between scores on certain standard tests and supervisory success in an aircraft factory. *J. appl. Psychol.*, 1946, 30, 328-332.
8. Selover, R. B. The development and validation of a battery of tests for the selection of clerical workers. *Amer. Psychologist*, 1948, 3, 291-292 - (abstract), and personal communication.

It is not intended to imply that these studies found highly predictive test batteries, but merely that the technique was adequate. Conclusive negative findings are important and are too frequently ignored or even suppressed.

In conclusion, let me emphasize two points. First, the actual work done by industrial psychologists is not as bad as would appear from this analysis, and the trend is definitely toward more complete and careful design and execution. In many cases our criticisms apply to the reports, not necessarily to the studies themselves. More care should be taken in the preparation of reports so that all relevant information is available to the reader.

Requirements of a "Good Report"

Perhaps a summary of the items one weary abstractor would like to see made explicit would be in order:

1. Detailed job description, with each group treated separately.
2. Complete description of the sample: N (sufficiently large), what proportion of the total population this represents and how selected, factors involved in hiring, age, length of time on the job (preferably with widely differing employees treated as separate groups), and total experience in jobs of similar nature; use of two samples, one an applicant group.
3. Exact test titles; when in the employment experience the tests were administered; whether the tests were a factor in hiring; where the tests were given; under what conditions and incentives the tests were given; reliabilities of tests with comparable groups.
4. Detailed description of the criterion; length of time on the job when the criterion measure was applied (with widely differing employees treated as separate groups); reliability of the criterion; some discussion of the validity of the criterion selected; if ratings are used, some estimate of the amount of contact the rater has with the employee; if production

records are used, the duration of the period and whether there were any unusual factors operating at that time.

5. Adequate statistical treatment, with assurance that the assumptions governing the use of the given measures have been met, and actual report of the numerical results, together with an appropriate measure of significance.

This may seem like a large order, but many adequately executed studies already reported could have included most of the items since it is obvious from certain remarks that the author must have taken them into consideration. In view of the untrustworthiness of many reports these items should be made explicit.

A final point concerns those studies done by inadequately trained personnel. There are many of these and they are quite useless. They point to the ultimate desirability of some method of identification of properly qualified personnel for employee selection programs.

Summary

A survey of more than 2,100 references on employee selection in industry has revealed that only 427 contain sufficient information to permit evaluation of the study. These 427 reports are analyzed in terms of annual volume, jobs most frequently investigated, statistics used in presentation of validity, number of subjects and general adequacy of design. This analysis reveals that many of these studies are inadequate to permit drawing conclusions as to the efficacy of the selection procedures

employed. Factors which influence results but are difficult to evaluate from reports as they are usually published are discussed. Some recommendations for items to be included in reports of employee selection programs are presented.

Received October 17, 1949.

References

1. Cronbach, L. J. Statistical methods applied to Rorschach scores. *Psychol. Bull.*, 1949, 46, 393-429.
2. Dorcus, R. M., and Jones, M. H. *Handbook of employee selection*. New York: McGraw-Hill, 1950.
3. Ghiselli, E. E. The validity of commonly employed occupational tests. *Univ. of Calif. Publ. in Psychol.*, 1949, 5 (9), 253-288.
4. Kurtz, A. K. A research test of the Rorschach Test. *Personnel Psychol.*, 1948, 1, 41-51.
5. MacMillan, M. H., and Rothe, H. F. Additional distributions of test scores of industrial employees and applicants. *J. appl. Psychol.*, 1948, 32, 270-274.
6. McMurry, R. N., and Johnson, D. L. Development of instruments for selecting and placing factory employees. *Advanc. Mgmt.*, 1945, 10, 113-120.
7. Rothe, H. F. Distribution of test scores of industrial employees and applicants. *J. appl. Psychol.*, 1947, 31, 480-483.
8. Rundquist, E. A., and Bittner, R. H. Using ratings to validate personnel instruments: a study in method. *Personnel Psychol.*, 1948, 1, 163-183.
9. Selover, R. B. The development and validation of a battery of tests for the selection of clerical workers. *Amer. Psychologist*, 1948, 3, 291-292, and personal communication.
10. Stromberg, E. L. Testing programs draw better applicants. *Personnel Psychol.*, 1948, 1, 21-29.

Cross Validation of an Abbreviated Point Job Evaluation System

Milton K. Davis and Joseph Tiffin

Occupational Research Center, Purdue University

Job evaluation has become widely used as a means of establishing equitable wage rates. Although there are many different methods available for constructing job evaluation systems, the most widely used technique involves the use of some type of a point scale. The approach used in point systems is to break the job into the various component items and assign points to each of these. The total points for all these items represent the evaluated job. During recent years there has been an attempt to construct scales which use fewer items than the longer and more involved ones.

Previous Studies. Lawshe and his associates (3, 4, 5, 6, 7) have published much material on such abbreviated scales. Chesler (1) has also published research on this subject. The procedure used in these studies was to select the three or four most important items from the longer system by the Wherry-Doolittle technique. Abbreviated scales thus derived were then compared with the original scales to determine the amount of agreement between the two. These studies have demonstrated that these abbreviated scales yield results which are comparable to the original system.

However, there has been some criticism against this approach since it was necessary first to analyze the original scale in order to derive the abbreviated one. Otis (8, p. 98) has stated this criticism in his recent book on job evaluation as follows: "How would one go about building an abbreviated scale? There is no way of knowing which one of the shorter scales obtained by Lawshe would best apply to a given plant. Constructing a shorter scale would necessitate a complete job evaluation using a longer scale. Either factor analysis or the Wherry-Doolittle technique would have to be applied to the data, and a shorter scale so derived would be used to keep the system up to date. These savings would not be very great."

Purpose of this Study. The present study was conducted to determine whether an ab-

breivated job evaluation scale, constructed in the light of key items previously identified by Lawshe (3, 4, 5), will achieve the same basic result as longer job evaluation systems. This abbreviated job evaluation system was derived without making a Wherry-Doolittle or factor analysis of the specific job evaluation data involved. Thus, the present study represents a cross-validation of this abbreviated job evaluation scale against the hold-out group of other job evaluation installations.

Procedure

Derivation of the Abbreviated Scale. The abbreviated point scale used in the present study was derived from the results of earlier research by Lawshe (4). That study analyzed the NEMA job evaluation system as it operated in three industrial plants. The NEMA job evaluation system, which consists of eleven items, was adapted by Kress (2) from the Western Electric Company's procedure for use by the National Electric Manufacturer's Association. It is one of the most widely used point job evaluation systems.

From the three multiple regression equations obtained in Lawshe's study, an average multiple regression equation was determined for cross-validation. Only those items which appeared in at least two of the three regression equations were used. These items were experience, unavoidable hazards, and initiative and ingenuity. Experience was the only item which was present in all three equations. The average value for each of these three items plus the average value for the constant was selected as the basis for further study. The final equation for predicting total points on the original system on the basis of the three items mentioned was:

Total points = $41 + 1.5$ (experience points) + 1.7 (initiative and ingenuity points) + 3.8 (hazards points).

Eight different companies submitted point

job evaluation data to Purdue University. Each of the sets of job evaluation data had been obtained with a system that included the three items mentioned above among a considerably larger number of items. All of the jobs represented were for hourly-paid industrial workers. Three of these installations were NEMA plans and the remaining five were other types of point rating systems. The total points for the abbreviated scale were calculated for each job from the equation given above.

Analysis and Results

Three types of analysis of the data were made. The first was a correlational analysis. Table 1 presents the correlations for each company between the total points from the regression equation on the three items and the original total points. Examination of this table shows that for those companies using the NEMA system the correlations were all .94 or above. Since these results were obtained from a multiple regression equation, the correlations presented correspond to multiple correlations between the abbreviated scale and the total point scale. It should be kept in mind, however, that since the multiple regression equation used came from previous studies of other plant installations, the correlations obtained by no means represent the maximum values that would have been obtained if the present data had been used in determining the multiple regression equation. The standard errors of the correlations are so small that any possibility of these representing chance fluctuations

from zero is virtually eliminated. Therefore, these correlations represent the extent of agreement between a single abbreviated point scale and the NEMA scale as it operated in three separate companies.

The second part of Table 1 shows the correlations between the total points from the regression equation on the three items and the original total points for the non-NEMA systems studied. In view of the method used to derive this abbreviated scale, its application to the latter companies might be open to question. However, these correlations were determined to find how well this one shortened scale operates for point scales which do not strictly follow the NEMA classifications and weights.

All the non-NEMA company plans yielded correlations above .90 with one exception. Company No. 5, a power and light utility firm, used a system which placed more weight on the unavoidable hazards item than is true with the NEMA scale. This fact probably accounts for the lower correlation.

Thus, the correlational analysis would seem to indicate that this abbreviated point scale has validity not only for the NEMA system but also for other point systems which are similar to the NEMA system and which include the items in the present abbreviated point scale.

Another method of analysis was in terms of labor grade displacement. This analysis was made in two separate ways, the first of which determined labor grade displacement when the new total point values were determined from the regression equation previously discussed, and the size, number, and limits of the labor grades for the simplified scale were exactly the same as those on the original scale. Therefore, the labor grades of the company scale represent the criterion against which the abbreviated scale labor grades were compared. This analysis could be made only for the first six companies mentioned in Table 1 because labor grades were not available for companies No. 5 and 6.

These results are summarized in Table 2. If the same or next adjacent labor grade is set as an acceptable standard, it can be seen that for the three NEMA installations 97 per cent, 96 per cent, and 94.5 per cent of the jobs meet this criterion. In company No. 2 there are 1.6 per cent of the jobs, representing only

Table 1
Correlations Between Three-Item Total and
Total Points

Company	No. of Jobs	<i>r</i>	S.E. _r
No. 1 NEMA	341	.96	.054
No. 2 NEMA	126	.95	.089
No. 3 NEMA	605	.94	.041
Non-NEMA Systems			
No. 4 (12 items)	61	.98	.129
No. 5 (11 items)	77	.69	.115
No. 6 (10 items)	185	.96	.074
No. 7 (11 items)	273	.96	.061
No. 8 (23 items)	253	.91	.063

two jobs, which would be displaced by three labor grades.

These results compare favorably with those previously reported by Lawshe (4). In plant A of his study, 99.2 per cent of the jobs would meet the criterion of the same or next adjacent labor grade.

For the three companies with the non-NEMA point ratings scales, the labor grade displacement was much greater. Table 2 shows that the percentages of jobs in the same or next adjacent labor grades for the three non-NEMA systems studied were, respectively, 42.6, 18.2, and zero.

Such low correspondence might seem to invalidate the use of the abbreviated scale with non-NEMA systems. However, a final analysis was made which overcame much of this difficulty. This analysis involved a determination of the range of points from the lowest to the highest evaluated job on the abbreviated scale. This range was then divided into the same number and with the same relative width of labor grades that existed for the same jobs on the original scale. Also, the lowest evaluated job on the abbreviated scale represented the lower limit of the first labor grade for the abbreviated scale. This method gave a consistent and reproducible approach for setting these labor grades.

Table 3 shows the results of this analysis. The percentage of jobs falling in the same labor grade, displaced one, two, and three or more labor grades is shown.

Table 2

Labor Grade Displacement: Based on the Abbreviated Point Scale with Number and Limits of Labor Grades the Same as in the Original System

Company	Per Cent of Jobs			
	Same Labor Grade	Displaced One Labor Grade	Displaced Two Labor Grades	Displaced Three or More Labor Grades
No. 1 NEMA	45.0	52.0	3.0	
No. 2 NEMA	52.5	43.5	2.4	1.6
No. 3 NEMA	44.6	49.9	5.5	
No. 4	9.8	32.8	37.8	19.6
No. 5	6.5	11.7	35.0	46.8
No. 6				100.0

Table 3

Labor Grade Displacement: Based on Proper Number of Labor Grades within the Abbreviated Point Range

Company	Per Cent of Jobs			
	Same Labor Grade	Displaced One Labor Grade	Displaced Two Labor Grades	Displaced Three Labor Grades
No. 1 NEMA	68.6	29.9	1.5	
No. 2 NEMA	56.3	39.7	2.4	1.6
No. 3 NEMA	44.8	49.4	5.8	
No. 4	42.8	47.6	9.6	
No. 5	31.2	49.3	19.5	
No. 6	36.8	49.2	13.5	.5

Table 3 demonstrates that for the three NEMA system companies 98.5 per cent, 96 per cent, and 94.2 per cent of the jobs fall into the same or next adjacent labor grade when the abbreviated point range method is used. However, these results do not differ materially from those obtained by the method which used for the abbreviated system labor grade classification the same size, number, and limits of labor grades as those used in the original system.

For the three non-NEMA systems there is definite improvement in accuracy of job placement by labor grade when the second method for placement is used. The respective percentages were 90.4, 80.5, and 86 for displacement of less than two labor grades. In addition, only one job in company No. 6 would be displaced by three labor grades.

The results thus show that the agreement between the labor grade placement of jobs by the long and the short scale is essentially the same when the original size and limits of labor grades are used for the short scale as when a proper division of the abbreviated point scale in the labor grades is used with this scale. For the non-NEMA systems, the proper allocation of labor grades within the abbreviated point range appears to result in greater agreement between labor grade placement by the long and short scales.

Summary

An abbreviated point job evaluation scale was constructed on the basis of prior research

on the NEMA job evaluation system. This abbreviated scale was based upon the average of three multiple regression equations found in that earlier study.

Eight job evaluation installations were studied for cross-validation of this one abbreviated scale. These installations included three NEMA systems and five non-NEMA point systems. All the jobs represented by these plans were hourly-paid.

Zero-order correlations between the total points of the abbreviated scale and the longer original scale were calculated for each of the eight installations. The amount of labor grade displacement was found by two methods. The first method compared the labor grade displacement when the labor grades for the abbreviated system were set up using the same limits as the original installation. In the second method for comparison of labor grade displacement, the abbreviated scale labor grades were set up by dividing the abbreviated scale range into the same number of labor grades as the original scale used for those jobs.

On the basis of this study, the following conclusions are supported:

1. For the NEMA installations, the correlations between abbreviated scale total points and the original scale total points were .96, .95, and .94.

2. This abbreviated scale will operate nearly as effectively for non-NEMA point job evaluation systems. With one exception, the obtained correlations of total points between the simplified scale and the non-NEMA point systems were above .90. A main requirement in applying this abbreviated scale to non-NEMA systems is that these other systems have items which closely approximate the three chosen items.

3. Labor grade displacement results show that 97 per cent, 96 per cent, and 94.5 per cent of the jobs in the NEMA installations would remain in the same or next adjacent labor grades when predicted from the abbreviated scale. These results do not change materially

when labor grades are set up using the abbreviated point range.

4. For the non-NEMA point systems studied, the superior approach was to divide the abbreviated scale point range into labor grades, the number and relative width of which are determined by the number and relative width of labor grades in the original scale. This method yielded 90.4 per cent, 80.5 per cent, and 86 per cent in the same or next adjacent labor grade.

5. From the results of the present study, it would seem that an abbreviated scale made up of three items—experience, hazards, and initiative and ingenuity—usually will achieve essentially the same results as a more extensive point system which includes these items.

Received October 1, 1949.

References

1. Chesler, D. J. Abbreviated job evaluation systems derived on the basis of "internal" and "external" criteria. *J. appl. Psychol.*, 1949, **33**, 151-157.
2. Kress, A. L. How to rate jobs and men. *Fact. Mgmt.*, 1939, **10**, 60-65.
3. Lawshe, C. H., Jr., and Satter, G. A. Studies in job evaluation. 1. Factor analysis of point ratings for hourly-paid jobs in three industrial plants. *J. appl. Psychol.*, 1944, **28**, 189-198.
4. Lawshe, C. H., Jr. Studies in job evaluation. 2. The adequacy of abbreviated point ratings for hourly-paid jobs in three industrial plants. *J. appl. Psychol.*, 1945, **29**, 177-184.
5. Lawshe, C. H., Jr., and Maleski, A. A. Studies in job evaluation. 3. An analysis of point ratings for salary paid jobs in an industrial plant. *J. appl. Psychol.*, 1946, **30**, 117-128.
6. Lawshe, C. H., Jr., and Alessi, S. L. Studies in job evaluation. 4. Analysis of another point rating scale for hourly-paid jobs and the adequacy of an abbreviated point scale. *J. appl. Psychol.*, 1946, **30**, 310-319.
7. Lawshe, C. H., Jr., and Wilson, R. F. Studies in job evaluation. 5. An analysis of the factor comparison system as it functions in a paper mill. *J. appl. Psychol.*, 1946, **35**, 426-434.
8. Otis, J. L., and Leukart, R. H. *Job evaluation*. New York: Prentice-Hall, 1948.
9. Peters, C. C., and Van Voorhis, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill Book Co., 1940.

Age and Route Sales Efficiency *

C. B. Cover and S. L. Pressey

Ohio State University

In the past fifty years, the average length of life in this country has increased 18 years, from about 49 to 67; and the proportion of people 45 and over has grown from 18% to 29% (4, pp. 50 and 256). Also, "there has also been a tendency over a long period of time to lower the age at which workers retire from active employment." And pressure for pension plans has accentuated a "wide-scale prejudice against hiring of the worker over 45 (2, pp. 63 and 20). The plight of the older workers thus mounts at the same time that their number increases.

Fortunately, the situation is receiving increasing attention, as evidenced by mentions in newspapers and magazines, and the efforts of such groups as the Desmond committee of the New York legislature (2, 3). Consideration has naturally tended to center on large-scale industry. But other common types of work may present certain of these problems in acute form. An occupation which initially gave such good financial returns that it tended to hold men into middle life past the age of ready occupational shift, gave little by way of experience which might help in any shift, but then dropped them or became so burdensome that they dropped out, would seem to present such problems. If thereby the business lost older workers whose experience might in some way have been utilized, it also suffered. This paper attempts briefly to outline such a situation.

Cases and Materials

The present study deals with 92 men whose work consisted of house-to-house selling of foodstuffs from a truck, making collections, and canvassing for new customers. Each salesman operated two routes on alternate days and thus served each customer three times a week. Pay, on a commission basis, was excellent,—a hundred dollars or more a week. A successful

salesman needed to have a pleasing personality, skill in selling, and enterprise in soliciting new customers. Presumably, all this involved knowledge of his goods, of demand in different neighborhoods, and of the likes of each customer. Since many of the sales were on credit, the man needed to be shrewd in his appraisal of customers; he needed also to estimate closely goods needed on each run so that there would be enough, but little surplus which might spoil or be salable only at a discount. He had also to be careful and efficient in his handling of his truck. Finally, the route salesman needed the physical stamina required for repeatedly getting in and out of his truck and going up and down steps carrying a basket loaded with merchandise, eight or nine hours a day, six days a week.

The findings are based on rank order ratings by sales supervisors of these men in: (1) total sales; (2) credit; (3) surplus; (4) truck repair and accidents; and (5) time. Ratings were based largely upon records and were thus fairly objective.

Results

The important features of the findings are exhibited in the following tables. In Table 1, the number at each age, as shown by the row of figures next to the bottom, should first be noted. Most men were in their twenties and thirties, with few staying beyond these years and only one (included with those in their fifties) past 60. Moreover, the first part of Table 1 shows that the older men who did continue fell off markedly in the basic issue of amount of sales. The twenty best salesmen are all under 40. Median rank drops markedly from the thirties to the fifties. A similar falling off was found in efficiency in use of time.

In contrast, the second part of Table 1 shows increasing efficiency with age in truck operation, as shown by repair costs and accidents, up through the forties; the twenty-two drivers who were poorest in these respects were all under 40. There was a similar improvement

* The data were gathered by Mr. Cover while a student at Ohio State University; he is now on the staff of Ohio University but in September, 1950, will become Assistant Dean of Students at Muskingum College.

with age in handling of credit, while judgment regarding surplus remained about the same. In short, though judgment appears to improve with age, a falling off in energy appears to cause a decrease in effectiveness in route salesmen.

The question naturally arises as to whether, if years of service instead of simply age were considered, ratings would be different. Table 2 so groups the cases, and in terms of ratings for total efficiency. Here the poorer ratings of the older salesmen are even more clearly exhibited! The two who have been salesmen longest are among the poorest.

The total task involved in these route sales jobs thus seems to be such that men do not maintain their efficiency past early middle life and most drop out. What happens to them? The truck sales positions pay very well and they have high status. Few of the men are willing to drop back to a much more poorly paying sales position in one of the company's stores. And as the business is organized, there are few supervisory or production jobs into which these men can go. As a result, though intimately familiar with the product sold and with sales problems, and potentially of continuing use to the firm, they commonly leave the business in their 40's. Some of them

Table 2

Total Efficiency of 92 Salesmen in Relation to Years of Service *

Rank Order	Years of Service			
	0-9	10-19	20-29	30-39
1-10	8	2		
11-20	9	1		
21-30	9	1		
31-40	9		1	
41-50	8	1	1	
51-60	8	1	1	
61-70	8		2	
71-80	9		1	
81-90	5	2	1	2
91-100	2			
Total	75	8	7	2
Median	43	36	68	86

* Median rank order for each length of service group is shown by N in bold-face type.

thereafter may do well. But most of them, so far as known, find no other satisfactory position at that age. And they become dissatisfied subsistence farmers, insurance salesmen, handymen, or otherwise on the margin of things vocationally.

Discussion

The data are indeed inadequate. Having been obtained after the war, they are undoubtedly influenced in various ways by circumstances related thereto. About the later careers of all too large a portion of the salesmen who had left, nothing was known. But known outcomes were so predominantly unsatisfactory that they seemed to point up certain problems of management and employee welfare. Might truck sales work be in some way made less fatiguing for older men, so that they might continue longer? Might the concern find more positions in its own organization into which these men might shift? Could the men's attitudes regarding the status of indoor selling jobs be so modified that they would be more willing to take them? Or could employment offices find more positions into which men with such experience could go, when they are in their forties or fifties? Situations of the type described in this brief paper seem not economically healthy, and in need of continuing study.

Table 1

Rank Order in Total Sales and Truck Operation and Accidents of 92 Route Salesmen, as Related to Age *

Rank Order	Age Groups: Total Sales				Age and Truck Operation			
	20-9	30-9	40-9	50-9	20-9	30-9	40-9	50-9
1-10	3	7			1	2	5	2
11-20	3	7			1	6	2	1
21-30	3	4	3		2	6	1	1
31-40	1	6	3		2	7	1	
41-50	4	4	1	1	4	2	1	3
51-60	5	4	1		3	5	2	
61-70	3	3	2	2	5	3	1	1
71-80	3	4	2	1	5	5		
81-90	4	3	1	2	5	5		
91-100				2	1	1		
Total	29	42	13	8	29	42	13	8
Median	52	36	46	80	64	41	18	36

* Median rank for each age group is shown by N in bold-face type.

Summary

With the increasing length of life and number of older workers in this country, it becomes increasingly important to investigate the relations of efficiency in different types of work to age, and to consider means by which individuals who are in occupations not feasible for the older years may find other vocational opportunities.

1. A study of 92 men selling foodstuffs from retail trucks showed increased efficiency with age in handling their trucks and in judgment in business relations with their customers (as in handling of credit), but a falling off in sales. The drain on physical energies involved in the work appeared to be the chief factor.

2. These men tended to leave such positions after 40; but most did not find other opportunities in the firm nor (so far as known) locate satisfying employment elsewhere.

3. The question is raised as to whether or not such work might be adapted so that older men could continue longer or other positions in the same firm found for them, which would utilize their experience. If not, might employment offices find opportunities for such men which, to some degree, would take account of their vocational background, and be more suited to their age and needs?

Received May 22, 1950.

Early publication.

References

1. Clague, E. After 45,—How about a job? *Survey Graphic*, 1950, 86, 173-176.
2. Desmond, T. C., and others. *Birthdays don't count*. New York State Joint Legislative Committee on Problems of the Aging, Newburgh, N. Y., 1948.
3. Desmond, T. C., and others. *Never too old*. New York State Joint Legislative Committee on Problems of the Aging, Newburgh, N. Y., 1949.
4. Dublin, L. I., and others. *Length of life* (Rev. Ed.). New York: Ronald Press, 1949.

Ortho-Rater Norms and Sex Differences

J. H. Ely, N. C. Kephart, and Joseph Tiffin

Occupational Research Center, Purdue University

The growing use of the Ortho-Rater as an industrial vision testing instrument has suggested the need for male and female industrial norms on the vision tests included in this instrument. It is the purpose of the present article to present such norms, based on quite

Table 1
Percentile Norms on Ortho-Rater Distance Acuity Tests

Score	Male				Female			
	Both Eyes	Right Eye	Left Eye	Worse Eye	Both Eyes	Right Eye	Left Eye	Worse Eye
0	1	1	1	3	1	1	1	2
1	1	2	2	4	1	2	2	3
2	1	3	3	5	1	2	3	5
3	1	3	4	6	1	4	4	7
4	1	4	5	8	2	5	6	10
5	2	6	6	10	3	8	9	13
6	3	8	9	14	6	12	14	20
7	8	12	14	20	16	20	22	29
8	14	22	22	31	27	37	38	48
9	25	34	33	45	44	53	54	66
10	37	43	44	56	62	67	69	79
11	59	61	66	76	82	82	86	92
12	87	92	91	98	95	97	97	99
13	96	99	97	99	98	99	99	100
14	98	99	99	100	99	100	100	100
15	100	100	100	100	100	100	100	100
N	7,659	7,657	7,648	7,646	2,469	2,469	2,467	2,468

Table 2
Percentile Norms on Ortho-Rater Near Acuity Tests

Score	Male				Female			
	Both Eyes	Right Eye	Left Eye	Worse Eye	Both Eyes	Right Eye	Left Eye	Worse Eye
0	1	2	3	4	1	1	1	2
1	1	3	3	6	1	2	2	4
2	1	4	4	7	1	2	3	5
3	2	6	6	9	1	3	4	6
4	4	7	8	12	1	4	5	8
5	6	9	10	14	2	5	6	9
6	9	15	15	22	4	9	11	15
7	16	20	23	29	10	15	17	23
8	26	29	31	39	23	28	28	38
9	46	49	46	60	48	59	53	70
10	65	66	59	73	74	79	71	85
11	87	80	77	87	93	92	89	95
12	98	95	95	98	99	99	98	99
13	99	99	99	100	100	100	100	100
14	100	100	100	100	100	100	100	100
15	100	100	100	100	100	100	100	100
N	7,655	7,652	7,641	7,637	2,468	2,469	2,468	2,467

Table 3

Percentile Norms on Ortho-Rater Vertical Phoria Tests

	Score	Male		Female	
		Far	Near	Far	Near
Left					
Hyperphoria	1	1	3	1	3
	2	2	7	2	7
	3	6	21	6	20
	4	23	55	24	56
	5	62	85	62	86
	6	87	95	86	95
	7	96	97	94	96
	8	98	98	96	98
Right					
Hyperphoria	9	100	100	100	100
	N	7,444	7,356	2,436	2,439

Table 5

Percentile Norms on Ortho-Rater Stereopsis and Color Tests

Stereopsis Test Score	Male	Female	Color Test Score	Male	Female
0	10	18	0	1	1
1	15	26	1	1	1
2	22	35	2	5	6
3	31	47	3	9	14
4	40	58	4	25	39
5	48	67	5	52	72
6	62	79	6	100	100
7	71	85			
8	79	89	N	7,597	2,457
9	100	100			
N	7,648	2,464			

large samples of industrial employees. A second purpose is to compare the mean scores and variability of men and women selected randomly from plants currently using the Bausch and Lomb Industrial Vision Service.

Norms

Tables 1 to 5, inclusive, give percentile norms on the several Ortho-Rater tests. These

Table 4

Percentile Norms on Ortho-Rater Lateral Phoria Tests

	Score	Male		Female	
		Far	Near	Far	Near
Esophoria	1	3	3	4	3
	2	6	6	7	6
	3	9	10	10	12
	4	15	18	15	22
	5	24	30	23	36
	6	37	45	34	53
	7	54	60	49	67
	8	73	72	68	77
	9	87	79	84	83
	10	94	85	92	87
	11	96	89	96	90
	12	98	92	97	93
	13	99	94	98	95
	14	100	96	99	97
Exophoria	15	100	100	100	100

norms are based on approximately 7,600 men and 2,500 women. At the bottom of each column in the tables is given the exact number of cases on which the individual test norm is based. The values in the tables are percentiles.

Table 6 summarizes the means, S.D.'s, and differences between males and females for the several tests in the Ortho-Rater. Quite a number of the differences, both in means and S.D.'s, are significant at the 5% level or below. Since the number of cases was large, several of the differences which are actually very small, were found to be significant at the 5% or even 1% level. Such differences are probably of more theoretical than practical importance.

It will be noted that in the color vision test a difference in favor of the men was found. This difference was significant at the 1% level. The authors are aware that this finding is contrary to long accepted theories and facts about the distribution of color blindness among the sexes. The explanation for this difference in findings is not known.

It should be kept in mind that while the data of this study were based on randomly selected men and women employed on industrial jobs, they may not be typical of randomly selected men and women from the general population. Traditional industrial practices of placing men and women on different types of jobs may have

Table 6

Comparison of Male and Female Means and Standard Deviation on Ortho-Rater Vision Tests
 Note: Numbers of cases in various categories shown at bottom of Tables 1 to 5.

	Male	Female	Differ- ence	C.R.	S.D.-male	S.D.-female	Differ- ence	C.R.
Far Acuity, Both	10.69	9.64	1.05	21.88*	2.07	2.07	.00	.00
Far Acuity, Right	10.11	9.12	.99	16.80*	2.63	2.52	.11	2.62*
Far Acuity, Left	10.06	8.97	.09	1.50	2.70	2.57	.13	3.10*
Far Acuity, Worse	9.26	8.27	.99	15.47*	2.94	2.69	.25	5.56*
Near Acuity, Both	9.43	9.47	-.04	1.00	2.12	1.60	.52	18.57*
Near Acuity, Right	9.17	9.03	.14	2.64*	2.77	2.12	.65	17.58*
Near Acuity, Left	9.22	9.12	.10	1.72**	2.94	2.35	.59	14.39*
Near Acuity, Worse	8.41	8.41	.00	.00	3.09	2.46	.63	14.65*
Far lateral phoria	7.07	7.25	-.18	2.90*	2.56	2.71	-.15	3.41*
Far vertical phoria	5.26	5.28	-.02	.63	1.28	1.38	-.10	4.35*
Near lateral phoria	7.22	6.80	.42	5.60*	3.23	3.14	.09	1.70**
Near vertical phoria	4.38	4.40	-.02	.87	1.37	1.40	-.03	1.30
Color	5.08	4.68	.40	14.81*	1.14	1.16	-.02	1.05
Stereopsis	5.23	3.97	1.26	18.53*	2.98	2.92	.06	1.25

* Significant at 1% level.

** Significant at 5% level.

resulted in certain secondary selection factors which would operate to make any randomly selected group of industrial employees (particu-

larly women employees) different in certain respects from the general population.

Received October 1, 1949.

Fluorescent Light Versus Daylight

J. Stanley Gray and Paul Prevetta

University of Georgia

Fluorescent lighting is comparatively new and few studies have been made of its effects on human vision. Luckiesh and Moss (1) compared fluorescent lights with tungsten lights of 20 foot-candles intensity by having 10 subjects read for 30-minute periods on 9 different days. The number of involuntary eye blinks during the first and last five minutes of each reading period was counted on the assumption that the frequency of blinks is related to the difficulty of seeing and the state of visual fatigue. They found that the increase in blinks was approximately the same for fluorescent lights as for tungsten lights. The small differences in increased blinking were not statistically significant.

However, Tinker (4) has criticized the rate of blinking as being a non-valid index of the visual function. He had 60 university students read under 10 foot-candles of well-diffused light and found that "the frequency of blinking is an unsatisfactory criterion of the readability of print." McNally (3) studied the relation of blink rate to reading various sizes of type and found no correlation. The rate of blinking was unrelated to the size of type. McFarland, Holway, and Hurvich (2) found an inconsistent relationship between the rate of blinking and both the duration of reading and the intensity of illumination. They concluded that blinking is not a reliable index of visual fatigue.

The authors of the present study used the American Optical Company's *Sight-screener* to measure the effects of two hours of continuous reading under daylight as compared with two hours of continuous reading under fluorescent lights. The visual functions measured were acuity, stereopsis, and both lateral and vertical phoria at 14 inches and at 20 feet distances. Fifty subjects (high school, college, and older adults) read books set in 8 point type for two hours under daylight of 20 foot-candles intensity and for two hours under fluorescent lights of 20 foot-candles intensity. For both reading

sessions, the light source was behind and above; books were kept on an easel at right angles to the line of vision; intensity was maintained at 20 foot-candles (venetian blinds were used to regulate the daylight); glare was eliminated by painting all peripheral background pastel green; the same book was read under both conditions; reading booths reduced outside distractions; and the visual skills were measured (using the *Sight-screener*) at the beginning, at the end of one hour of reading, and at the end of two hours of reading for both sessions. Some of the subjects read first under fluorescent lighting and then under daylight; and others read first under daylight and then under fluorescent lights.

There was both loss and gain in all four visual skills at both near and far measurements. Some subjects actually increased in certain visual skills after the two-hour reading period although most subjects lost in all skills except stereopsis. The summarized results are shown in Table 1. The differences between

Table 1
Visual Effects of Two Hours of Reading under Daylight and Fluorescent Light for Fifty Subjects

	Daylight (Mean Change)	Fluorescent (Mean Change)
Near Distance (14 inches)		
Acuity*	-.04	-.12
Stereopsis*	.08	.00
Lateral Phoria†	-.31	-.02
Vertical Phoria†	-.01	.05
Far Distance (20 inches)		
Acuity*	-.04	.02
Stereopsis*	-.20	-.20
Lateral Phoria†	-.26	-.21
Vertical Phoria†	-.02	.11

* Data in "units of change."

† Data in diopters.

daylight and fluorescent light were very small and none was statistically significant.

Summary

The conclusion seems to be justified that, as measured by the sight screener, fluorescent light of 20 foot-candles intensity is not inferior to daylight of the same intensity for reading 8 point type material for two hours duration.

Received May 22, 1950.
Early publication.

References

1. Luckiesh, M., and Moss, F. K. Vision and seeing under light from fluorescent lamps. *Illum. Engng.*, N. Y., 1942, 37, 81-88.
2. McFarland, R. A., et al. *Studies in visual fatigue*. Cambridge, Mass.: Harvard School of Business Administration, 1942.
3. McNally, H. J. The readability of certain type sizes. *Teach. Coll. Contr. Educ.*, 1943, No. 883.
4. Tinker, M. A. Validity of frequency of blinking as a criterion of readability. *J. exper. Psychol.*, 1946, 36, 453-460.

Inconsistency in the Predictive Value of a Battery of Tests

Robert M. W. Travers

Division of Teacher Education, Board of Higher Education, New York City

and

Wimburn L. Wallace

University of Massachusetts

The increased number of applicants for admission to professional schools in recent years has made administrators keenly aware of problems of selection. A growing dissatisfaction has been felt with the traditional criteria for selection, such as undergraduate grades or letters of recommendation, and some attention has been devoted to the possibility of using tests for selective purposes. Various large-scale testing programs have been initiated for this purpose in many professional areas, while in others, smaller programs have been developed for studying the problem of selection. The schools of dentistry have shown admirable and exemplary caution in this matter, and while the Association of Schools of Dentistry has sponsored an experimental testing program, it has not introduced the tests as a requirement for admission, prior to adequate investigation. The use of tests for the selection of students of dentistry is still largely experimental.

During 1947 and 1948, the University of Michigan has administered tests to applicants for admission to its School of Dentistry. The primary purpose of this battery was to explore possible improvements in the admission system. The tests were selected on the basis of previous studies which now have been adequately reviewed in a bulletin published by the Veterans Administration¹ and which need not be reviewed again here. The one exception in this procedure was the inclusion of a test of "Effectiveness of Expression" which, it was believed, might measure a competency important to the dentist though possibly of little importance in achieving satisfactory grades in dental school.

The following tests were included in the bat-

tery: ACE Psychological Examination, 1945 Edition; MacQuarrie Test for Mechanical Ability; Bennett Test of Mechanical Comprehension, Form BB; Revised Minnesota Paper Form Board Test, Series MB; Interpretation of Reading Materials in the Natural Sciences, College Level, Tests of General Educational Development; and Effectiveness of Expression, Cooperative English Test B2, Form T.

The tests were administered at the University of Michigan and at centers throughout Michigan. A few centers were also established outside of the State. In the Spring, 1947 testing program there were 32 centers, but when the program was given again in the spring of 1948 these centers were reduced to 18. At each center the appointed examiner was given a set of directions to be read *ad verbatim* to the examinees.

The scores on the tests were correlated with the subsequent achievement of those admitted. These correlation coefficients are given in Table 1.

This table shows an apparent inconsistency: the tests had some predictive value when they were given to the group admitted in 1948, but practically no predictive value when they were given to the previous class. A number of hypotheses suggest themselves concerning this inconsistency, the most obvious one being that the tests were improperly administered during 1947 but were properly administered in 1948. This hypothesis is easily tested since a group of tests, similar in many respects to the tests discussed, was administered to all students after admission through a testing program sponsored by the American Dental Association. If the tests given in 1947 were improperly administered one could expect that they would have low validity, but that the battery administered by the American Dental

¹ *Predicting success in training for dentistry.* Veterans Administration Technical Bulletin TB 7-44, July 8, 1947, pp. 16.

Table 1
Coefficients of Correlation Between Honor Point Ratio and Test Scores

Tests	1947 Admissions (N = 82)		1948 Admissions (N = 79)
	1st Year Honor Point Ratio	2nd Year Honor Point Ratio	1st Year Honor Point Ratio
ACE Quantitative Score.....	-.02	-.08	.47
ACE Linguistic Score.....	-.06	-.09	.39
ACE Total Score.....	-.05	-.10	.48
MacQuarrie.....	-.05	-.10	.27
Mechanical Comprehension.....	.14	.07	.28
Paper Form Board.....	.06	.05	.25
GED Natural Sciences.....	.20	.19	.45
Effectiveness of Expression.....	.01	-.02	.36

Association would not be affected by this factor and should have validity comparable with that of the tests given in 1948.

The correlations of the tests in the American Dental Association battery with honor point ratio are summarized in Table 2.

The tests of the American Dental Association show the same phenomenon exhibited by the University of Michigan tests. They have little predictive value for those admitted in 1947 but substantial value in 1948. Hence, it is unlikely that improper administration of the 1947 battery can account for the phenomenon.

The next hypothesis concerning the change in the predictive value of the tests from 1947 to 1948 is that the 1947 group exhibited a more limited range of ability. Table 3 shows the means and standard deviations of the scores of the two groups.

The data presented do not support the hypothesis that the lack of predictive value of the tests given in 1947 is due to a restricted range of test scores or of achievement (HPR).

A third hypothesis is that the system of grading changed from 1947 to 1948. This hypothesis does not seem to be consistent with the evidence. A discussion of the problem with administrative officials indicated that no changes had been made in the grading system.

A fourth hypothesis is that the group selected in 1947 was in some way different from that selected in 1948. Differences between the mean scores of the two groups are small and substantially the same part of the measuring scale was used for prediction in the two cases. It is, however, possible that in the case of the group admitted in 1947, there might have been a tendency on the part of the admissions officer

Table 2
Coefficients of Correlation Between American Dental Association Tests and Honor Point Ratio

Tests	1947 Admissions		1948 Admissions
	1st Year Honor Point Ratio	2nd Year Honor Point Ratio	1st Year Honor Point Ratio
ACE Quantitative Score.....	.09	.05	.35
ACE Linguistic Score.....	.09	.07	.27
ACE Total Score.....	.05	.00	.41
½ GED Reading.....	.18	.18	.34
GED Spelling.....	-.03	-.03	.42
English,—Total.....	.05	.04	.47
Object Visualization.....	.25	.23	.25
Natural Science Survey.....	.26	.29	.44

Table 3

Standard Deviation and Mean of Test Scores of Those Admitted in 1947 and 1948

	1947 Admissions		1948 Admissions	
	Mean	Standard Deviation	Mean	Standard Deviation
ACE Quantitative Score	51.6	8.4	50.5	7.94
ACE Linguistic Score	75.2	11.8	78.2	12.94
ACE Total Score	136.9	17.05	128.7	18.23
MacQuarrie	78.8	12.25	78.0	10.46
Mechanical Comprehension	41.7	8.73	37.0	10.87
Paper Form Board	47.6	6.69	48.3	7.25
GED Natural Sciences	56.8	10.82	50.3	11.00
Effectiveness of Expression	46.4	8.77	43.6	10.36
Honor Point Ratio (1st Yr.)	1.56	0.44	1.53	0.45

to accept *either* high grades in pre-professional work *or* high test scores for admission. If this were the case to any extent it might result in the selection of a group in which there was a correlation of zero between previous scholastic record and test scores. This seems to be the case, for as Table 4 demonstrates, the group selected in 1947 shows a generally lower correlation between previous scholastic record and test scores than those selected in 1948.

Table 4

Coefficients of Correlation Between Test Scores and Average Grade in Pre-professional Work

	1947 Admissions	1948 Admissions
ACE Quantitative Score	.22	.28
ACE Linguistic Score	.11	.25
ACE Total Score	.19	.30
MacQuarrie	-.29	.02
Mechanical Comprehension	-.18	.24
Paper Form Board	.03	.16
GED Natural Sciences	.22	.37
Effectiveness of Expression	.07	.25

If the type of selective process under discussion were operative in 1947 one would expect not only that the group selected would show exceptionally low correlations between test scores and previous scholastic record but also that there might be exceptionally low correlations between test scores and subsequent performance.

Summary

This study has been presented to illustrate the fact that the process of validating tests for admission to an educational program by finding the correlation with subsequent grades needs to be more carefully scrutinized. It is possible that the correlations between test scores and grades may be zero although the instrument is valid for selection purposes. The process of selection may control the size of validity coefficients and should be properly controlled in any experimental validation of a test.

Received October 7, 1949.

Intercorrelations in Merit Rating Traits

C. E. Jurgensen

Minneapolis Gas Company

Rating scales used by different companies vary considerably in the number of traits rated. In an analysis of 132 rating scales, Mahler (6) found that the number of traits varied from one to thirty-three with a mean of 9.3. Companies using such scales appear to make two assumptions: (1) the traits rated are important from management's point of view, and (2) ratings on the various traits are discrete. This study deals only with the second of these assumptions.

Kornhauser (5) reported intercorrelation of instructors' ratings on 68 students who were rated on intelligence, industry, accuracy, cooperativeness, initiative, moral trustworthiness, and leadership. Intercorrelations ranged from .45 to .83 with a median of .69.

Driver (2) reported a study of a ten trait merit rating scale from which several traits had been discarded because ratings were not discrete. The intercorrelations of the remaining ten traits (based on N's varying from 100 to 300) ranged from .11 to .79 with a mean of .46.

Ewart, Seashore and Tiffin (3) reported a study of ratings of 1120 men on a twelve trait scale covering safety, knowledge of job, versatility, accuracy, productivity, overall job performance, industriousness, initiative, judgment, cooperation, personality, and health. Intercorrelations on the twelve trait scale ranged from .25 to .88 with a median of .75. Reliability coefficients were not reported. A factor analysis showed that a general factor (ability to do the present job) accounted for most of the total variance. A second, and oblique, factor was tentatively named "skill possessed over and above the requirements for the specific job." The third and final factor (health) was discarded as an artifact.

Bolanovich (1) made a factor analysis of ratings on 143 field engineers whose work required a wide range of abilities and characteristics. They were rated on fourteen traits: personality, personal appearance, punctuality,

thoroughness, efficiency, resourcefulness, dependability, cooperation, job attitude, technical ability, sales ability, organizing ability, judgment, and desire for self improvement. Intercorrelations ranged from .05 to .73 with a median of .49. A factor analysis resulted in six common factors: attendance to detail, ability to do the present job, sales ability, conscientiousness, organizing or systematic tendency, and social intelligence. The correlation between factors was not reported.

Intercorrelations given in the above studies appear to be typical of those usually reported when rating scales consist of relatively narrow and specifically defined traits. Such narrowness is generally believed to be advisable. For example, Paterson (7) has said "each trait to be rated should be restricted to a single type of activity or to the results of a single type of activity, otherwise the ratings will be ambiguous." Although much can be said in favor of restricted traits, the fact remains that intercorrelations have generally been high. The hypothesis was therefore proposed that relatively independent factors would be obtained if each factor consisted of a cluster of sub-traits which might logically be expected to be highly correlated in the rating situation, and if the major factors could logically be expected to be relatively discrete in actuality. The testing of this hypothesis was a primary purpose of the research reported here.

The hypothesis was also proposed that reported intercorrelations between traits are spuriously high due to differences between raters in leniency and variability. The possible effect of such differences seems to have been given consideration with respect to interpretation of ratings on specific employees, but to have been ignored consistently with respect to intercorrelation of traits. The testing of this hypothesis was also a primary purpose of the research reported here.

Description of Scale

The rating scale developed for this study consisted of six traits which were thought to be reasonably discrete plus ratings on overall value and potentiality. The nature of the scale is illustrated by the first trait:

Work Habits. Consider initiative, industry, persistence, accuracy, carefulness of detail, orderliness, thoroughness, speed, punctuality and other related work habits.

- 0—Unsatisfactory, does not meet job requirements
- 1
- 2—Below average, partially meets job requirements
- 3
- 4—Average, meets job requirements
- 5
- 6—Above average, exceeds job requirements
- 7
- 8—Outstanding, far exceeds job requirements

Definitions of other traits used in the scale are given below. Descriptions of degrees were similar in nature to those given under *Work Habits*.

Attitudes. Consider employee's enthusiasm, interest, loyalty, willingness to cooperate, teamwork, open mindedness, tolerance, ability to accept criticism, and other such attitudes toward the job, fellow workers, supervisor, and company.

Acceptance by Others. Consider whether this employee is liked and respected by co-workers and public, how they react to his appearance, voice and manner, tact, courtesy, agreeableness, friendliness, etc.

Self Control. Consider his emotional maturity and balance, disposition, stability, calmness, maturity of action, etc.

Mental Ability. Consider his alertness, reasoning ability, common sense, memory, judgment, analytical ability, ability to express self orally or in writing, mechanical intelligence, ease of learning new work and remembering instructions, and other similar aspects of intelligence.

Physical Ability. Consider his agility, coordination, dexterity, vision, hearing, endurance, energy, strength, overall health, etc.

Overall Rating. Decide whether he is currently an asset or liability to your department, considering all evidence (whether mentioned in above characteristics or not) which affects his all-around performance at the *present* time.

Potentiality. To what extent does this employee possess capacity for future growth? Consider his all-around performance at the present time in light of his age, health, experience, etc., and the extent to which these and

other factors will foster or prevent future development.

Method

Instructions to raters required them to decide to what degree the employee possessed the trait under consideration and to encircle the number which best described him with respect to the position he held. Ratings on all traits were on a nine point scale, each description including one or two equivalent phrases. Descriptions of degrees were selected on the basis of a rough application of Thurstone's method for scaling equal appearing intervals (4).

Results reported here are based on ratings of 199 employees by 26 supervisors. The employees were all hired within a six week period, were all on the same job, and were randomly divided into 26 crews having seven or eight employees each. The work was such that supervisors had considerable contact with their crew members on an individual basis, but relatively little contact with their crew as a unit. Ratings were obtained three months after employment and again five months after employment. Reliability coefficients for the eight traits were obtained from the repeated ratings.

Each supervisor had rated seven or eight men on two occasions on eight traits. Therefore approximately 120 ratings by each supervisor were available. Inspection of these ratings showed that supervisors differed considerably both in mean rating and variability of ratings. On the nine point rating scale, mean ratings by supervisors ranged from 3.7 to 7.0. Variability, in terms of sigma, ranged from .8 to 2.1. All ratings were converted to standard scores (having a mean of 50 and sigma of 10) on the assumption that differences in ratings were due to differences between raters rather than employees. The further assumption was made that a rater who differed from others in leniency or variability on one trait would differ similarly on other traits. These assumptions appeared tenable in this situation.

Results

Intercorrelations of ratings expressed in standard score form are given in Table 1. They range from .33 to .84 with a median of

Table 1

Reliability and Intercorrelations of Ratings-Standard Score Form (N = 199 Conversion Employees)

	Work Habits	Atti- tudes	Acceptance by Others	Self Control	Mental Ability	Physical Ability	Overall Rating	Potenti- ality
Mean	48.1	49.9	50.2	48.4	49.3	49.1	47.7	46.6
Sigma	10.1	10.9	9.2	9.3	8.6	8.7	11.0	11.0
Work Habits	.55							
Attitudes	.66	.54						
Acceptance by Others	.41	.33	.45					
Self Control	.68	.68	.53	.56				
Mental Ability	.65	.60	.48	.63	.53			
Physical Ability	.53	.52	.45	.63	.55	.53		
Overall Rating	.84	.62	.52	.71	.66	.59	.67	
Potentiality	.63	.55	.49	.65	.74	.51	.58	.68
Median	.65	.60	.48	.65	.63	.53	.62	.58

.60. These same intercorrelations expressed in original raw score ratings range from .60 to .88 with a median of .76. Each of the 28 intercorrelations dropped in size when ratings were expressed in standard score form, the range of drop being from .01 to .44 with a median of .16.

Reliabilities of ratings in standard score form are also given in Table 1. They range from .45 to .68 with a median of .55. Reliabilities computed from original raw score ratings range from .68 to .75 with a median of .73. Each of the eight reliabilities dropped in size when ratings were expressed in standard score form, the range in drop being from .07 to .25 with a median of .18.

Summary

The attempt to develop a six trait scale in which each trait consisted of a cluster of sub-traits did not result in a scale wherein the major traits were independent of each other. Reliabilities and intercorrelations of these traits were essentially of the magnitude usually reported for more restricted traits. The first hypothesis was therefore refuted.

Inasmuch as intercorrelations between narrowly restricted traits tend to be high (often as high as their reliabilities) and inasmuch as the same situation has been found to exist with regard to broadly worded factors consisting of clustered sub-traits; it would therefore appear unnecessary and unwise in many rating situations to obtain trait ratings. This would be the case when trait ratings (weighted or un-

weighted) are summed to obtain a composite score which is to be used for any purpose. It is simpler, more direct, and equally effective to obtain an overall rating instead of a composite based on highly correlated trait ratings. This does not necessarily deny all value of trait ratings. Overall ratings may be more valid or reliable if made after consideration has been given to traits, even though trait intercorrelations are high. Second, even though intercorrelations are high, there will be some few individuals who are rated high on one trait and low on another. In such cases the trait ratings can be helpful. This appears particularly true when a primary purpose of ratings is to serve as a basis for a supervisor-employee conference on employee progress.

When converted to standard score form, intercorrelations dropped in magnitude but reliabilities dropped equally as much. Therefore the various trait ratings remained of little if any value as discrete traits. The consistent drop in reliabilities which occurred when original ratings were converted to standard scores indicates that reliability coefficients may be spuriously high if obtained from ratings not equated for differences in mean and sigma from one rater to another. The second hypothesis was therefore confirmed. Unless an investigator gives evidence that he has considered and eliminated the effects of any such differences, his reliability coefficients should be questioned to the point of rejection.

Received October 24, 1949.

References

1. Bolanovich, O. J. Statistical analysis of an industrial rating chart. *J. appl. Psychol.*, 1946, 30, 23-31.
2. Driver, R. S. A case history in merit rating. *Personnel*, 1940, 16, 137-162.
3. Ewart, E., Seashore, S. E., and Tiffin, J. A factor analysis of an industrial merit rating scale. *J. appl. Psychol.*, 1941, 25, 481-486.
4. Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1936, 440-444.
5. Kornhauser, A. W. A comparison of ratings on different traits. *Personnel J.*, 1927, 5, 440-446.
6. Mahler, W. R. Some common errors in employee merit rating practices. *Personnel J.*, 1947, 26, 68-74.
7. Paterson, D. G. Principles of merit rating. *Personnel Digest*, 1944, 1, 16-20.

How Readable Are Corporate Annual Reports? *

Siroon Pashalian

New York University

and

William J. E. Crissy

Queens College

Business enterprises have long been concerned with communication problems. Today there is increasing interest in how best to "get the word around" to jobholders, shareholders, customers, and the general public. One of the first formal media of communication to be used was the corporate annual report. In recent years an extensive literature has developed concerning how best to construct and publish such reports (1, 2, 7, 8). Most of the papers have reflected judgments of varying degrees of expertness rather than findings based upon experimental research.

A current problem in the construction of the annual report is that of endeavoring to make it more understandable and more widely read. Readership surveys show a considerable apathy to company reports. Yet, it is by no means an easy task to prepare adequate and concise reports. The challenge is one of presenting sufficient technical data and information for the financial expert, satisfying the requirements of the law (especially in railroad and public utility reporting), and at the same time, being meaningful to those whose interests are of a more general nature. The present demand crystallizes as the need for writing an informative account of the year's operations, and for presenting the material in such a way that the report will be read.

In connection with this problem, one of the writers undertook to investigate the readability

* This paper is a report on some of the findings from Pashalian's M.A. thesis entitled, "An Investigation of the Application of the New Flesch Readability Formulas to Corporate Annual Reports," submitted to the Department of Psychology, Graduate School of Arts and Science at New York University, October 1949. Explorative findings concerned with the relationship between Flesch indices and judges' ratings of readability have not been reported due to basic design limitations.

of corporate annual reports by means of the new "Flesch Readability Formulas" (5, 5a).

Method

The annual reports of those corporations that are listed in the Corporate Billion-Dollar Club in the June 11, 1949 issue of *Business Week* were included in the present study. These members are either non-financial companies with assets of over \$1-billion, or with annual revenues or sales of over \$1-billion, or both. Presumably then, these corporations have the largest number of stockholders, employees and other persons interested in their operations. In other words, there is substantial public interest in these big corporations; their annual reports are expected to reach a vast audience.

Applying the sampling technique suggested by Flesch (3, 4, 5), one-hundred-word samples were chosen from every other page of each of these twenty-six reports. This procedure, with no restrictions on the number of samples to be taken per report, is believed to have achieved a fair sampling in proportion to the length and breadth of the report. A total of 211 samples were examined; the average number of samples per report was 8.1.

Results

The findings on the application of the new Flesch Readability Formulas to the twenty-six annual reports are listed in Table 1.

Analysis of the Reading Ease Measures

The range of readability scores was from 6 to 58. According to Flesch reference categories for these scores (5, 5a), these reports vary within descriptive styles of *very difficult*,

as with material of scientific and professional journals, to *fairly difficult*, as in literary and quality magazines, such as *Harper's*. This range, interpreted in terms of the educational attainments of the U. S. adult population, suggests a potential audience of from 4½% of the population completing college, to 40% of the population who have had some high school education (4).

The average Reading Ease score for the entire set of reports was 34.37. Writing at this level is generally *difficult*, and descriptive of the style in academic material; for example, the *Yale Review*, which may be comprehended by 24% of the population who have graduated from high school or have had some college training.

The measure of average sentence length

ranged from 16—*fairly easy*, typical of slick-fiction and understandable by 80% of the population, to 53—*very difficult*, above much scientific material and understood by approximately 4½% of the population.

The measure of the average number of syllables per 100 words ranged from 156—*fairly difficult*, to 183—*difficult*. Flesch has advised that a comfortable text contains one and one-half times as many syllables as words (6). The preponderance of reports in the *difficult* category on this measure (22 of the 26 reports) is indicative of a high level of abstraction in the language of these reports.

In relation to this syllable measure, the factor of numbers was encountered in the material of the annual reports. Under Flesch's directions, numbers separated by space are

Table 1

Summary of Reading Ease and Human Interest Measures on the Twenty-six Annual Reports

Industry and Average R.E.	Corporate Annual Report	Average Sentence Length	Average Number Syllables	Reading Ease Score	Percentage of Personal Words Sent.		Human Interest Score
Merchandise	Sears, Roebuck	22	162	47	2	0	7
44.50	Montgomery Ward	16	175	43	0	0	0
Communications	Bell Telephone	24	165	43	1	0	4
43.00							
Foods	Swift & Co.	17	156	58	5	2	19
43.00	Armour & Co.	21	168	43	4	0	15
	Safeway Stores	27	182	28	2	0	7
Autos & Accessories	General Motors	21	174	38	2	0	7
40.00	Chrysler Corp.	20	171	42	1	0	4
Oil	Standard Oil (N. J.)	21	179	34	0	0	0
33.50	Standard Oil (Ind.)	25	173	35	2	0	7
	Socony-Vacuum	21	180	33	2	0	7
	Texas Co.	23	182	32	1	0	4
	Gulf Oil	24	175	34	1	0	4
	Standard Oil (Cal.)	24	178	32	0	0	0
Utilities	Consolidated Edison	27	166	39	1	0	4
31.50	Commonwealth-Southern	31	183	24	0	0	0
Railroads	Pennsylvania	27	170	36	1	0	4
28.50	NY Central	25	174	34	1	0	4
	Southern Pacific	32	175	26	0	0	0
	Santa Fe	32	177	25	1	0	4
	Baltimore & Ohio	18	171	44	2	0	7
	Union Pacific	53	174	6	0	0	0
Machinery & Supplies	General Electric	30	178	26	0	0	0
26.00							
Metals & Chemicals	U. S. Steel	32	173	28	0	0	0
25.67	E. I. du Pont	31	176	26	1	0	4
	Bethlehem Steel	39	172	23	1	0	4

counted as words in any text; several and lengthy figures should be omitted from the syllable count. Instead, a corresponding number of words to the number of figures omitted should be added, and their syllable totals added to those already counted (5). When applying these directions to the samples, the number of figures per 100 words was also recorded. The average number of figures per 100 words ranged from 2.30 to 8.80. The highest number of figures per 100 words ranged from 5 to 21. Thus, a surprisingly large number of figures appeared in small samples of 100 words. Their disastrous effect on the general reader not widely trained in numbers or mathematics invites speculation. It would seem, therefore, that greater care and attention should be given to determining best ways of presenting figures in such reports. Hundred-word samples that are crowded with 10 to 20 lengthy figures should caution writers or editors, and suggest their more effective incorporation in a table or chart.

The significance of these results on readability obtained by analysis utilizing the Flesch technique is perhaps best indicated by showing what would improve the "scores" made. A need is demonstrated for more effective use of punctuation devices. The semicolon, for instance, can be more widely used to shorten sentence lengths and at the same time, to maintain any indications that the words and information belong closely together.

Similarly, the need for writing at a less difficult level is indicated. A "writing down" would not necessarily be an under-estimation of intelligence. People whom corporations want to influence probably range from low normal intelligence to the superior. They have the capacity to grasp such concepts as gross sales, profits, etc. However, the vast audience which these reports reach is assumed by corporate reporters to possess far greater language facility than it does. The language of these reports, shown in terms of the education of the U. S. adult population above, is too difficult for the great bulk of this diversified readership to comprehend. Corporate writers are over-estimating the language experience of their potential audience—stockholders, employees and the general public.

Human Interest Value

The range of Human Interest scores was from 0 to 19. Again, in terms of Flesch references (5, 5a), these styles are from *dull*, descriptive of the style in scientific journals, to *mildly interesting*, descriptive of trade magazines. The average Human Interest score for the entire set of reports was 4.27—*dull*.

Thus, the corporate writing of these twenty-six annual reports is extremely low in human interest value, i.e., in the personal words and sentences which provoke and continue general reader interest, and help the reader to understand the text better. In an era which is fostering the teamwork and cooperation of stockholders and employees alike, the need for the stress on "we" and "our" and "us" is inescapable. "I" can serve to bridge the tremendous gap between the President and stockholders and employees. This set of reports used such words sparsely. Personal words help to convey the feeling in the material of having been written directly to the reader, whoever he may be. They can reflect the whole spirit and tone of the organization.

In addition, corporate writing can direct greater attention to individual personalities. Although the sampling necessarily tapped only certain pages, extremely few samples mentioned specific persons and their accomplishments. It appeared that much of this kind of information was confined to Employee Relations headings or obituaries. People are interested in people, and they want to become better acquainted with the outstanding personalities of the corporation. Yet, among the 21,100 words sampled in this study, only approximately twenty names were mentioned, and even these were noticeably concentrated within certain reports.

Moreover, to enhance human interest value, there remains the need for the appropriate use of personal sentences—exclamations, questions and commands directly expressed to the reader. Only one report among the twenty-six possessed a sentence of this description in the sampling scheme—a question. More question-marks in annual reports can provoke thought, continue or revive reader interest. Similarly, direct commands are another interest-controlling device. Instead of the impersonal, "It will be

noted . . . ,” “Note . . . ” can do much more to invoke the effort of a glance at the charts and an independent analysis.

Sample Passages

To illustrate the various levels of difficulty obtained by means of the new Flesch Readability Formulas, sample passages from the annual reports of those corporations which ranked twenty-sixth, thirteenth, and first in Reading Ease are furnished below:

From the Union Pacific Railroad report:

Capital Stock

At the annual meeting of Union Pacific Railroad Company stockholders held on May 11, 1948, the Articles of Association were amended so that on July 1, 1948 the total number of authorized shares of preferred and common stocks of the Company were doubled (with no increase in the total aggregate par value thereof), and the then outstanding 995,431 shares of \$100 par value preferred stock became 1,990,862 shares of \$50 par value preferred stock, and the then outstanding 2,222,910 shares of \$100 par value common stock became 4,445,820 shares of \$50 par value common stock, each of the new \$50 par value preferred and common shares being entitled to one vote at any meeting of stockholders.

From the New York Central Railroad report:

Dieselization is progressing

Carrying forward our motive power modernization, the Central and leased lines, together with two affiliates, the Pittsburgh & Lake Erie and the Indiana Harbor Belt Railroads, ordered in 1948 new Diesel-electric locomotives at a total cost of approximately \$33,600,000. The bulk of these locomotives, on which deliveries will extend into 1950, are for road freight and for switching service. The Central's portion was about \$24,790,000.

Locomotives delivered during 1948 increased the Dieselized portion of the total road freight train mileage of the Central and leased lines to approximately 13.5 per cent by the end of the year.

From the Swift and Company report:

What Swift & Co. is Trying To Do

The public rightly expects a business to accomplish certain desirable things.

Who determines what is desirable in a free country? Not one man or a group of men. Each individual decides for himself whether he will buy from, sell to, work for, or invest in a company.

A decision to buy a product is a vote in its favor. The votes of millions of people may cause prices to go up or down. The results quickly tell what the public thinks desirable.

Such economic democracy can thrive only in a certain climate—one in which prices are free and competitive, and business is spurred by the hope of profits and the fear of losses.

Analysis by Industries

The arrangement of the twenty-six reports by industry in Table I facilitates interesting and noteworthy comparisons. There seems to be a certain amount of homogeneity within industries on all the obtained measures of the Flesch formulas. At the same time, however, it must be remembered that the entire set of reports has demonstrated a great degree of homogeneity and narrow range under the Flesch technique.

The reports of railroad companies have the greatest amount of variability on the measures employed. Their range in average sentence length is from 18 words (*standard*) to 53 words (*very difficult*). This observation seems to reflect the great difficulty attached to railroad reporting due in part to legal specifications concerning content. Apparently, however, some railroad companies are fulfilling their legal and public obligations in a more effective manner of readability than others.

Another striking inference may be obtained from Table 1. The arrangement by rank order in readability attainments under the formulas, parallels almost directly the degree of contact the companies have with the general public. Merchandising, Communications, Foods, Automobiles and Accessories corporations cater to larger sections of the general public. Corporations dealing in Machinery and Supplies, Metals and Chemicals have a more restricted, less diversified market for their products (8). Since this observation is based on the small and variable groups of the study, however, more data and analysis are actually required for final proof.

Nevertheless, such an arrangement is by no means unwarranted. It is generally accepted that the extent and character of the public interest should first be determined in the construction of the report. Then a corporate writer attempts to write to that audience. However, if this same trend had appeared

within lower degrees of difficulty, it could be considered a more legitimate consequence of the nature of the enterprise and the groups interested in its operations. At the same time, it would then meet the readability requirements of these particular audiences.

Summary

1. Analysis of the readability of the twenty-six annual reports of corporations listed in the Billion-Dollar Club of *Business Week*, June 11, 1949, by means of the new Flesch Readability Formulas, revealed that, on the whole, the general level of reading was *difficult*, and the human interest value *dull*.

2. These reports contain language which is beyond the language experience and fluent comprehension of approximately 75% of the U. S. adult population.

3. The Flesch technique demonstrates promise as a method for indicating the difficult language elements in corporate reports.

4. It also demonstrates promise as a method for spotting "impersonalness" in such writing.

When, as in this study, the writing sample involves a problem of mass communication, the Flesch technique appears to be a reasonable instrument. It gauges the likelihood that these annual reports will convey their messages to most of their prospective readers. Wider application of the technique in the construction

of the annual report is recommended. Used in conjunction with the other types of practical hints in the literature, it can serve to strengthen the annual report as the most important single written communication between management and stockholders, employees, and the general public.

Received October 7, 1949.

References

1. Dale, E. *Preparation of company annual reports*. Research Report No. 10, American Management Association, New York, 1946, 104 pp.
2. Doris, Lillian. *Modern corporate reports*. New York: Prentice-Hall, 1948.
3. Flesch, R. F. *Marks of readable style; A study in adult education*. New York: Bureau of Publications, Teachers College, Columbia University, 1943. (Contributions to Education, No. 897.)
4. Flesch, R. F. *The art of plain talk*. New York: Harper and Brothers, 1946.
5. Flesch, R. F. A new readability yardstick. *J. appl. Psychol.*, 1948, 32, 221-233.
- 5a. Flesch, R. F. *The art of readable writing*. New York: Harper and Brothers, 1949.
6. Flesch, R. F. Making the narrative readable. Chapter 15 in *Modern Corporate Reports* by Lillian Doris, New York: Prentice-Hall, 1948.
7. Gibson, W. B., ed. *The annual report; A study of over 500 financial reports of leading American Business Institutions showing the present style trend and important physical characteristics*. Chillicothe, Ohio: Mead Corporation, Marketing Research Division, 1939.
8. McLaren, N. L. *Annual reports to stockholders—Their preparation and interpretation*. New York: Ronald Press, 1947.

Rorschach Responses, Strong Blank Scales, and Job Satisfaction Among Policemen

Solis L. Kates

Michigan State College

An individual with certain personality characteristics, just short of thoroughgoing psychotic disorganization, may be attracted to and satisfied with an occupation because his personality traits are compatible with its demands. Certain personality traits, whether they contribute to a high degree of adjustment or maladjustment may be of importance for or lead to the development of interest and satisfaction in a particular occupation. For instance, many of the employed patients seen at a mental hygiene clinic gave Rorschach responses that were markedly abnormal yet they were able to carry on satisfactorily in their occupations.

The behavior of an individual should be studied not only in relation to the total culture but with regard to his sub-cultural group and its demands upon him. When the personality tendencies of an individual are compatible with the demands of a particular sub-culture, probably little anxiety or self-dissatisfaction may be provoked. If these personality traits conflict with the sub-cultural demands, probably much anxiety and self-dissatisfaction may ensue. Because occupational groups are readily available for study, they have been selected here as a particular type of sub-culture for scrutiny. This study proposes only to ascertain the personality traits that appear to be possessed by policemen. Many occupations will have to be investigated to ascertain the personality attributes associated with membership in the occupation. But occupational groups must be carefully selected and the subjects identified with respect to the homogeneity of their vocational interests. It is interesting to note that studies in the field of vocational choice and achievement have been suggested by Rapaport (9) as a method of validating the Rorschach Test.

The hypothesis that a significant relationship exists between measured vocational interests and job satisfaction postulated by

Strong (11) requires substantiation. An investigation recently completed by the writer (3) did not demonstrate any significant relationship between measured vocational interests and job satisfaction of routine office workers. More evidence is essential with respect to this hypothesis.

Subjects: The subjects of this study were twenty-five New York City Patrolmen who volunteered while off duty for testing. All came from one precinct. Their average age was 32.8 years and their mean educational attainment was 12.2 years. Due to the small sample, the results discussed below must be considered as suggestive rather than conclusive.

Method: The Rorschach Test was administered in groups of four or five and an individual inquiry was conducted with each subject. The subjects completed the Strong Vocational Interest Blank and a job satisfaction blank of the Hoppock type. The Rorschach responses were scored as suggested by Klopfer (4) and were evaluated by means of the Munroe Inspection List (6) which yielded a total score indicative of the degree of maladjustment. The subjects' responses were appraised according to the extent of their deviation from clinically established normal limits in terms of 28 Rorschach categories.

Results

The mean policemen interest score for the subjects was 40.6 and the standard deviation was 10.1. The difference between this mean interest score and that for Strong's criterion group was significant at the one per cent level. However, when the scores were divided into two categories corresponding to A and B+ ratings and those B and below in ratings, this distribution was not significantly different from the expected distribution of scores according to Strong's data (11).

Table 1

Product-Moment Intercorrelation Among Police Interests, Job Satisfaction, Occupational Level, and Munroe Inspection Technique Scores

	Police Interests	Job Satisfaction	Occupational Level	Munroe Scores
Police Interests	—	.35	-.74**	.19
Job Satisfaction	.35	—	-.51**	.47*
Occupational Level	-.74**	-.51**	—	-.26
Munroe Score	.19	.47*	-.26	—

* Significant at five per cent level.

** Significant at one per cent level.

The mean job satisfaction score of the policemen was 20.0, with a standard deviation of 3.2, while the mean Munroe Inspection score was 11.9 and a standard deviation of 3.7. The mean occupational level score on the Strong blank for the policemen was 48.9, standard deviation 5.7, which mean was not significantly different from that of Strong's criterion group.

Table 1 lists the product-moment correlation coefficients between the variables under consideration, namely, job satisfaction, occupational level, measured police interests, and Munroe Inspection scores.

The biserial correlation coefficients between some of the Rorschach response categories on the Munroe Inspection List and job satisfac-

tion, police interests, and occupational level are listed in Table 2.

In Table 3, the respective quantities of the whole (W) and human movement (M) responses are shown that have been arbitrarily considered as the optimum ratios. If the number of whole (W) responses was smaller than its lower limits for the specific number of human movement (M) responses given in the protocol, the ratio was considered to be excessively in favor of the human movement (M) responses. If the number of whole (W) responses was greater than its upper limits in relation to the human movement (M) responses then the ratio was considered to demonstrate an excessive preponderance of whole (W) responses.

Table 2

Biserial Correlation Coefficients Between Munroe Inspection List Categories and Police Interests, Job Satisfaction, and Occupational Level

Rorschach Score	Police Interests	Job Satisfaction	Occupational Level
Color: Movement	.53*	.03	-.47*
CF: FC	-.25	-.02	.12
FC	-.31	-.03	.61*
FM: M	.11	.00	.00
M	.61*	.00	-.52*
Total Color	.55*	.52*	-.58*
Fc	-.18	.00	-.25
F%	.25	.20	-.30
Popular	-.07	-.51*	.29
Dd	-.18	.29	.10
W: M	—	—	-.60*
W: M(M > W)	—	—	.62*

* Significant biserial correlation coefficients.

Discussion

The mean score of the subjects was significantly lower than that of Strong's criterion group, indicating that many subjects possessed

Table 3

Arbitrary Optimum Numbers of Whole (W) Responses in Relation to Human Movement (M) Responses

Numbers of Human Movement (M) Responses	Arbitrary Optimum Numbers of Whole (W) Responses
1	2 to 4
2	3 to 8
3	5 to 12
4	6 to 15
5	8 to 15
6	9 to 15
7	10 to 16
8	11 to 18

extremely low measured police interests. On the other hand, the fact that the proportion of A and B+ interest ratings was as expected in terms of Strong's criterion group illustrated that letter ratings were probably of greater accuracy than the magnitude of interest scores in forecasting the composition of individuals within an occupational group. Furthermore, letter ratings should be given more value than the magnitude of standard scores in guiding individuals. Finally, it may be concluded that many subjects entered police work because of their genuine interest while the others had different reasons.

The degree of job satisfaction was significantly greater than that of routine office clerks (3), significantly lower than that of nursing students (7), and slightly but insignificantly lower than that of engineering students (1). There was probably greater opportunity for policemen to express their abilities and predilections in the pursuit of their duties than existed for routine clerks. The factor of greater remuneration of the policemen was an important element that could not be evaluated.

The degree of maladjustment of the policemen, as measured by the Rorschach Test was slightly but insignificantly smaller than that of routine clerks (3) and slightly but insignificantly greater than that of biologists (10). Probably, the policemen, as a whole, demonstrated as many signs of maladjustment as may be found in other groups.

No significant relationship was found to exist between the degree of measured police interests and satisfaction with police work. Consequently, the police interest scale of the Strong Blank cannot be used with any degree of certainty in forecasting satisfaction with police work. One of the reasons for this absence of relationship may be assigned to the lack of precision of the measuring instruments. Moreover, Strong in setting up his occupational scales, did not attempt to differentiate the satisfied from the dissatisfied members of the occupation. It was unwarranted to assume that those with the higher interest scores would be more satisfied than those with the lower scores. Similarly, since there was no significant relationship between job satisfaction and police interest ratings when the latter were separated into A and B+, and B and lower

ratings, letter ratings may not be utilized as a predictor of work satisfaction.

Another reason for this absence of a significant relationship between job satisfaction and measured interests was the fact, as Strong indicated, that for some occupations, the weeding out process is not thoroughgoing and the elimination of candidates is not in terms of their personal attributes. Hence, the members of such occupations may not be homogeneous. What would satisfy one need not satisfy another member of the occupation. The more homogeneous the members, the more will similar occupational pursuits satisfy the members. The greater the homogeneity of members of an occupation, the greater and more significant will be the relationship between measured interests and job satisfaction. The less homogeneous the members, the smaller will be the relationship between measured interests and job satisfaction. Policemen may not be as homogeneous as, for example, physicians and engineers and hence the relationship between their measured interests and job satisfaction may not be as great as that between physicians and engineers. Furthermore, in vocations where personal skills are required and the tasks are challenging, it is probable that the correspondence of measured interests to those of other successful members of the occupation would be significantly associated with work satisfaction. [A more extensive treatment of this hypothesis can be found in the monograph by Kates (3).]

Finally, the unlikely possibility must be kept in mind that measured interests may change with experience on the job. Probably, the stability of measured interests goes hand in hand with the homogeneity of members of an occupation by being related as well to the extent and character of the elimination process. Where the weeding out process is extensive and is based on personal attributes, the measured interests of the members of the occupation should be quite stable. Where the elimination process is not so extensive and is not based on personal attributes, measured interests may be subject to change by further experience on the job.

The significant negative relationship between the occupational level and police interests is quite similar to that found by Strong. For

those policemen with high measured interests, the occupational level, probably indicative of vocational aspiration level (2), was low.

The significant negative relationship between job satisfaction and occupational level lent further support to the acceptance of the occupational level scale as an indicator of the level of aspiration in vocational endeavor. The policemen who had high occupational levels of aspiration were prone to be dissatisfied with their jobs. It illustrates the fact that in police work, the occupational level is a more potent contributor to job satisfaction and dissatisfaction than measured police interests. Probably, it may be stated that, for policemen, job satisfaction is not as intimately linked with measured police interests as it is related to their occupational levels of aspiration. The subjective reactions of the policemen in terms of their aspiration levels, as measured by the occupational level scores, were of greater importance to their job satisfaction than the mere possession of high police interests. The fact that the member may be similar in his measured interests to other successful members of his occupation may not be as important to his satisfaction as the fact that he has not attained to a level of vocational achievement that he believes necessary for his success. This conclusion is consistent with the experiments of Lewin (5) demonstrating that the level of aspiration is a more vital factor for satisfaction with achievement than is objective performance. Probably in the higher status occupations such as physician and lawyer, there would be a positive relationship between occupational level and job satisfaction.

The more maladjusted policemen, as measured by the Rorschach Test and Munroe Inspection Technique (assuming that the Munroe is a valid criterion of maladjustment for N.Y.C. policemen), tended to be more satisfied with their work than the less maladjusted. Hence, this finding corroborated the hypothesis that the individual's maladjustments might contribute to his job satisfaction rather than to his job dissatisfaction. It controverted the thesis that most job dissatisfaction stemmed from the individual's maladjustment.

It was ascertained that policemen with high measured police interests could be or could not be maladjusted. Similarly, high occupational

levels of aspiration were not associated with maladjustment. In police work, it may not be concluded that the individuals with the greater degree of vocational ambition tended to be more maladjusted.

High police interest scores were associated with the tendency of policemen to give significantly more movement (M, FM, m) responses than color responses (FC, CF, C) to the Rorschach cards. They were prone primarily to be motivated by and responsive to inner promptings rather than to stimuli from without.

High police interests were related to the tendency for policemen to give an adequate number of human movement (M) responses of reasonably accurate form. Low police interests were associated with the giving by policemen of poor human movement (M) responses or two and less human movement (M) responses. It may be said that the policemen who were prone to accept their inner promptings as constructive and positive forces would have high police interests.

The higher the police interests, the greater was the tendency for the subjects to give two and less color (FC, CF, C) responses. Policemen with high police interests did not possess the general readiness to establish emotional relationships with the world around them.

In addition, the higher the job satisfaction score, the greater the tendency of policemen to give two and less color (FC, CF, C) responses. Satisfied policemen were not prone to establish a ready emotional relationship with the world around them. It is probable that policemen who tend to establish good social relations would not be happy on their jobs. Subjects who as a matter of practice did not establish good social relations would probably find their conditions of work rewarding.

High job satisfaction was related to the giving by policemen of three and less popular responses. The more satisfied policemen were not likely to think along conventional lines.

No other significant relationships were discovered between the Munroe Inspection List categories and police interests and job satisfaction.

The higher the occupational level of aspiration, the greater the tendency for policemen to give total color (FC, CF, C) responses that were equal to or greater than total movement

(M, FM, m) responses. Higher occupational levels of aspirations were associated with the policemen's responsiveness to the environment about them. The same trend was evident in the finding that high occupational levels were related to the giving of an adequate number of form color (FC) responses. Rational emotional reactions to stimuli by policemen were associated with high occupational levels of aspiration. Finally, high occupational levels of aspiration were given by policemen who had an adequate number of color (FC, CF, C) responses demonstrating further that the ability to be responsive to environmental stimuli went hand in hand with high vocational aspiration. It does sound reasonable and valid to have policemen with high occupational levels of aspiration display characteristics that are recognized as necessary to achieve higher economic status. In another study, a definite relationship has been demonstrated to exist between great energy and activity in out-of-school and co-curricular activities of high school boys, and high occupational level scores (8). Since responsiveness to outside stimuli is self-evident as a trait for carrying on co-curricular and out-of-school activities, one of the results of this study is to furnish additional evidence validating the meaning of the Rorschach color responses.

Low occupational levels of aspiration were associated with adequate human movement (M) responses of policemen. High occupational levels were related to less adequate numbers of human movement (M) responses and poorly seen human movement (M) responses. The more nearly the ratio, whole (W) responses to human movement (M) responses, approached an arbitrary optimum, the lower the occupational level of policemen. When the number of human movement (M) responses in relation to whole (W) responses was in excess of the arbitrary optimum, the occupational level of the policemen tended to be high.

It may be concluded that when policemen had an adequate number of human movement (M) responses in terms of both the Rorschach protocol and whole (W) responses, they were inclined to accept their inner promptings and their present status in life. They were not prone to aspire to higher occupational levels as

they had probably accepted as suitable their own outlook on life as well as the realization that their attainments were consonant with their ability. Policemen with inadequate human movement (M) responses tended not to accept their status in life, and appeared to have a need to aspire to higher occupational levels. The same conclusion applied to a preponderance of human movement (M) responses in relation to whole (W) responses with the probable additional meaning that the excessive inner strivings within the policemen drove them toward higher occupational levels.

Summary

The subjects of this study, New York City policemen, were found: (a) to demonstrate no significant difference from Strong's criterion group in measured police interest ratings, (b) to have a relatively high job satisfaction level, and (c) to be no more maladjusted than routine office clerks and biologists.

No relationship was discovered between the subjects' measured interests and job satisfaction, between measured interests and Rorschach maladjustment, and between occupational level and Rorschach maladjustment. Significant relationships were found to exist between measured police interests and occupational level, between job satisfaction and occupational level, and between job satisfaction and Rorschach maladjustment. In these conclusions, the assumption was made that the Munroe Inspection Technique is a valid criterion of maladjustment for New York City policemen. Several reasons were advanced for the presence and absence of the above relationships.

Additional support has been received from the results of this study to validate the meaning of the color (FC, CF, C) responses. The human movement (M) responses have been assessed in a different context and additional meanings have been attached to the relations of the human movement (M) response to the whole (W) response and to the total Rorschach protocol.

Investigations of policemen who are undergoing treatment as a consequence of emotional maladjustment should be of help in ascertaining whether the hypothesis obtains, namely

that the personality traits of the particular sub-cultural group with which the patient was identified, should be known to evaluate his relative degree of maladjustment.

Finally, policemen with high police interests tended to be markedly introversive, to have adequate ability to accept their own strivings and outlook as mature, and to be relatively unresponsive to stimuli from without. Satisfied policemen tended to be relatively unresponsive to stimuli from without and to be lacking in the capacity to think along conventional lines. Further studies of the personality of policemen, if they confirm the above results, will be instrumental in helping to validate the significance of some Rorschach responses.

Received October 17, 1949.

References

1. Berdie, R. E. The prediction of college achievement and satisfaction. *J. appl. Psychol.*, 1941, 25, 197-204.
2. Darley, J. G. *Clinical aspects and interpretation of the Strong Vocational Interest Blank*. New York: The Psychological Corporation, 1941.
3. Kates, S. L. Rorschach responses related to vocational interests and job satisfaction. *Psychol. Monogr.*, 1950, 64, 3 (Whole No. 309).
4. Klopfer, B., and Kelly, D. *The Rorschach technique*. New York: World Book Co., 1942.
5. Lewin, K., Dembo, T., Festinger, L., and Sears, P. S. Level of aspiration. In J. McV. Hunt (Ed.) *Personality and the behavior disorders*. New York: Ronald Press, 1944, Vol. I, 333-378.
6. Munroe, Ruth L. The inspection technique: a method of rapid evaluation of the Rorschach protocol. *Rorschach Res. Exch.*, 1944, 8, 46-70.
7. Nahm, Helen. Satisfaction with nursing. *J. appl. Psychol.*, 1948, 32, 335-343.
8. Ostrom, S. R. The OL key of the Strong Test and drive at the twelfth grade level. *J. appl. Psychol.*, 1949, 33, 240-248.
9. Rapaport, D., Gill, M., and Shafer, R. *Diagnostic psychological testing*. Chicago: Year Book Publishers, Inc., 1946. Vol. II.
10. Roe, Anne. Analysis of group Rorschachs of biologists. *Rorschach Res. Exch.*, 1949, 13, 25-43.
11. Strong, E. K., Jr. *Vocational interests of men and women*. Stanford, California: Stanford U. Press, 1943.

Card Versus Booklet Forms of the MMPI

Wm. C. Cottle

The University of Kansas

The purpose of this article is to compare the card *versus* the booklet form of the Minnesota Multiphasic Personality Inventory (MMPI) to determine whether it would be possible to use these forms interchangeably in a testing program. Is it possible to consider them as comparable forms of the test?

Wiener (1), alternating the card and booklet form in chronological order with each new counselee at the St. Paul Guidance Center, found that "there are no differences between the individual and group forms on any scales that approach statistical significance."

Holzberg and Allessi (4) using 30 psychiatric patients tested within one or two days on the full and abbreviated form of the MMPI, report that "although statistically significant differences were found between the mean weighted scores of half the scales, these results were not clinically significant as judged by profile results."

The Sample

A brief description of the sample of 100 cases included here is given in Table 1. It will be noted that the educational level of the group is relatively high since all are college students, with approximately half the group undergraduates and the rest graduate students in the School of Education at the University of Kansas. Modal school grade attained by this group was first year of graduate study.

The group volunteered to take both forms of the test and no pressure was exerted to secure compliance.¹ As evidence of this only 100 of the proposed group of 175 completed both forms of the test within the one week time-limit imposed.

Age of the group ranged from 18 to 60 with a median age for the 68 males of 27.8 years and for the 32 females of 21.5 years.

¹ Acknowledgment is made to Professors E. T. Gaston, A. H. Turney, J. F. Nickerson, G. M. Carney, and J. O. Powell of the University of Kansas for making it possible for their students to take these tests.

An undergraduate group of music education students contained 23 males and 19 females. In a subgroup of music education teachers taking graduate work there were 15 males and four females. In a subgroup of miscellaneous teachers and public school administrators there were 30 males and nine females. These comprise the total group of 68 males and 32 females.

Table 1
The Sample

A.	Age	Male	Female
	18-20	7	13
	21-25	17	10
	26-30	18	2
	31-35	8	0
	36-40	10	3
	41-45	3	3
	46-50	4	0
	51-55	1	0
	56-60	0	1
	Total	68	32
B.	School Grade	Male	Female
	13	9	10
	14	1	
	15	13	9
	17	36	13
	18	7	
	19	2	
	Total	68	32
C.	Occupation	Male	Female
	Mus. Ed. Student	23	19
	Mus. Tchr.	15	4
	Misc. Tchr.	30	9
	Total	68	32
D.	Card Form Taken	Male	Female
	First	34	20
	Last	34	12
	Total	68	32

Table 2

Means and Standard Deviations of Card vs. Booklet Form of MMPI for 100 College Students

Scales		L	K	F	Hs	D	Hy	Pd	Mf	Pa	Pt	Sc	Ma
68 Males	$\bar{X}$	2.53	15.57	2.72	3.88	16.84	18.88	13.34	24.91	7.88	8.91	7.82	15.97
Card	σ	1.68	4.99	2.34	3.57	3.94	5.15	4.32	5.40	2.72	6.99	6.46	5.63
68 Males	$\bar{X}$	2.64	15.81	3.56	3.99	18.54	19.71	14.19	26.00	8.84	10.18	8.82	15.97
Booklet	σ	1.25	4.49	2.78	3.62	4.69	5.00	4.41	5.62	2.72	7.81	7.13	5.31
32 Females	$\bar{X}$	2.56	15.44	2.44	5.25	17.87	21.44	13.13	36.09	8.22	11.41	10.16	14.65
Card	σ	1.22	5.06	2.09	4.78	4.16	5.19	4.31	5.41	2.53	7.13	7.22	4.63
32 Females	$\bar{X}$	2.91	15.97	2.47	5.13	18.63	21.63	13.78	37.47	8.72	11.09	9.09	15.93
Booklet	σ	1.31	4.94	2.25	5.00	5.06	5.66	4.75	4.44	2.68	8.16	8.50	4.38
100 Cases	$\bar{X}$	2.54	15.53	2.63	4.32	17.17	19.70	13.33	28.49	7.99	9.71	8.57	15.07
Card	σ	1.55	5.01	2.27	4.06	4.04	5.30	4.32	7.50	2.67	7.13	6.80	5.02
100 Cases	$\bar{X}$	2.73	15.86	3.21	4.35	18.57	20.32	14.06	29.67	8.80	10.47	8.91	15.94
Booklet	σ	1.28	4.64	2.68	4.15	4.82	5.30	4.53	7.51	2.71	7.94	7.60	4.70

Procedure

It was intended that one-half of the males and females would take the card form first and the booklet form second within one week. The other half would take the booklet form first and the card form second within a week. The booklet form was administered in a group situation and the card form was given individually. In actuality, as shown in Part D of Table 1, the planned procedure was achieved with the male group, but not with the female group. Twenty females took the card form first and only 12 females took the card form last.

Raw scores without the "K" correction were recorded for each individual on each form and Pearson product moment correlation coefficients were computed for males, females and the entire group on each scale. These are shown in Table 3.

Means and standard deviations for the entire group and separate groups of 68 males and 32 females are shown in Table 2. The profile of mean scores for males is shown in Figure 1, and that for females is shown in Figure 2.

In order to examine the magnitude of the standard error of estimate where card and booklet forms are assumed to be equivalent

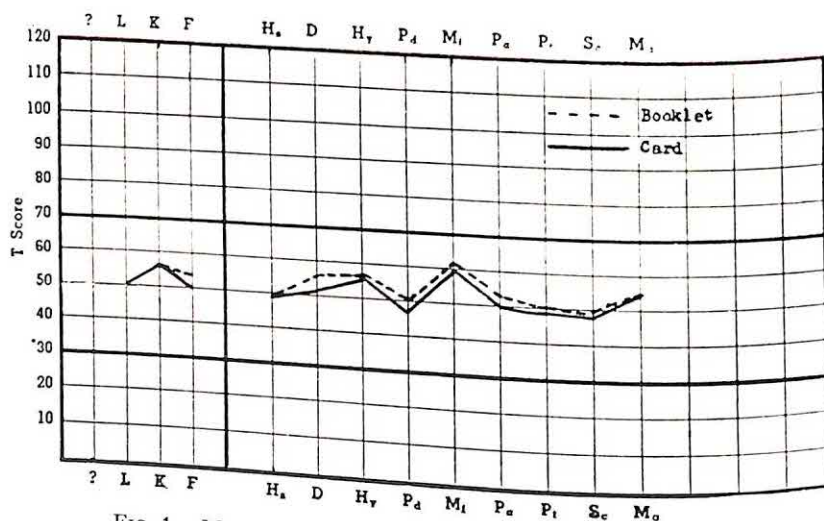


FIG. 1. Mean T-scores on the MMPI for 68 college males.

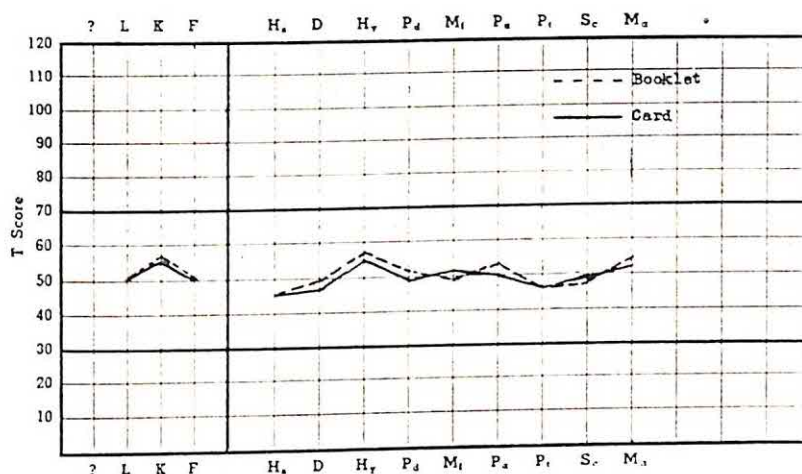


FIG. 2. Mean *T*-scores on the MMPI for 32 college females.

and to check the accuracy of the product moment correlations, a table of score deviations was constructed and correlation coefficients were computed by means of the ratio of the estimated true variance to the observed variance. In these computations variance error was computed as one-half the variance of the distribution of difference in raw score between the card and booklet form and observed variance was taken to be the average of the variance of the card and booklet form.²

Table 3 shows for each scale the correlation coefficient secured in this manner (r'_{cb}), the variance of the total distribution of differences in raw score between the two forms for each scale (V_d), the average of the variance for both forms (V_0), the standard error of estimate in raw score points, the approximate *T*-score change at one standard error, and the variance ratios of card and booklet forms.

² The formulae used for these computations were:

$$V_d = i^2 \left[\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N} \right)^2 \right]$$

Where V_d is variance of difference between raw scores on card and booklet; i is an interval of 1 raw score point; $N = 100$; and fx is frequency of deviations per step interval.

$$r'_{cb} = 1 - \frac{V_e}{V_0}$$

Where r'_{cb} is the correlation between card and booklet form; $V_e = V_d/2$, and V_0 is average variance of card and booklet form.

Results

With the exception of the coefficients shown for the lie scale (L), depression (D), and paranoia (Pa), the product moment correlation coefficients reported in Table 3 range from .72 to .91. Omitting the validity scales, those representing actual syndromes of maladjustment approach the size of accepted reliability coefficients. The most reliable scales would appear to be masculinity-femininity (Mf),³ psychasthenia (Pt), and schizophrenia (Sc); all three above .85.

Holzberg and Alessi (4) report test-retest coefficients on the short and long form of the MMPI taken within three days which range from .519 to .927. With the exception of hypochondriasis (Hs), psychopathic deviate (Pd), and hypomania (Ma), these coefficients ranged between .72 and .93. The scales above .85 were the lie scale (L), validity (F), hysteria (Hy), and schizophrenia (Sc).

Rotter, in commenting upon the test-retest reliability of the card form of the MMPI, reports coefficients ranging .71 to .83 (2). Hathaway in personal correspondence with this writer indicates that test-retest coefficients ranging from .61 to .85 have been secured on seven scales using individual *versus* group form.

³ Reference to Table 3 shows that the separate coefficients for males and females on the Mf scale are both lower than the coefficient for the entire group on this scale. Combining these groups increases the range and consequently the size of the correlation coefficient for the total.

Hathaway and McKinley (3) report reliability coefficients for the card form as follows:

D with 40 normal cases, $r = .77 \pm .044$
 Pt with 47 normal cases, $r = .74 \pm .15$ (test-retest)
 Pt with 200 normal cases, $r = .91 \pm .07$ (corr. split half)
 Hy with 47 normal cases, $r = .57$ (test-retest)
 Ma, number not given, $r = .83$ (test-retest)
 Pd with 47 normal cases, $r = .71$ (test-retest)

Capwell (5) reports test-retest correlations on 85 public school girls ranging from .40 to .77, only three of which are above .72. She reports similar correlations on 98 home school girls which range from .33 to .71. The difficulty in comparing those studies to the present study is that only the one by Holzberg and Alessi employs a similar short time between tests. That is, it is difficult to estimate how much the reliability coefficient is affected by personality changes occurring in the length of time between tests.

It would appear that, with the exceptions noted above for L, D, and Pa, correlation coefficients between card and booklet forms of the MMPI reported here in Table 3 are as high as those reliability coefficients reported for the card form alone. This would suggest that the group or booklet form of the MMPI could be used interchangeably with the card form for a college group. Saving in time required to administer and score the booklet form make this preferable in many psychometric situations.

Wiener (1), reports that on seven of the

scales there is a slightly higher average score on the individual form than on the group form. Reference to Table 2 indicates that males score slightly higher on the booklet (group) form for all scales, except Ma, where mean scores are identical.

Mean scores for females are slightly higher for the booklet form also, except for Hs, Pt, and Sc. The combined mean scores for the 100 cases ($P < .01$) are all slightly higher for the booklet form. It would appear that there is a tendency for a college population to score slightly higher on the booklet than on the card form of the MMPI. A college population tends to place fewer items in the "Cannot Say" category when the booklet form is used.

These results do not necessarily conflict with the research of Wiener, because his group was a non-disabled male veteran population in a guidance center operated by the Department of Education at St. Paul, Minnesota.

Let us consider the results shown in Table 3. It is evident there is no significant difference in the result of the two methods of computing the correlation coefficients between card and

booklet form. The $\sqrt{\frac{V_d}{2}}$ is the best estimate of the standard error of the estimate, e.g., one-half the variance of the raw score difference between card and booklet forms on the Sc scale is 7.76 and the standard error of the estimate is calculated as 2.78. This means

Table 3
Correlation Coefficients and Related Estimates for Card vs. Booklet Form
of the MMPI with 100 College Students

	N	L	K	F	Hs	D	Hy	Pd	Mf	Pa	Pt	Sc	Ma
r_{cb} Males	68	.51	.79	.77	.72	.65	.65	.81	.83	.53	.92	.90	.78
r_{cb} Females	32	.34	.72	.72	.91	.69	.83	.79	.79	.63	.87	.82	.75
r_{cb} (Product Moment)	100	.46	.76	.75	.81	.66	.72	.80	.91	.56	.90	.86	.76
$r'_{cb} = 1 - V_e/V_0$	100	.46	.78	.74	.81	.65	.73	.80	.91	.57	.88	.85	.76
$V_e = V_d/2^*$	100	1.15	5.09	1.61	3.26	6.92	7.77	4.06	5.19	3.16	6.64	7.76	5.64
$V_0 = V_e + V_b/2^*$	100	2.05	23.57	6.23	17.05	20.01	28.38	19.85	56.97	7.34	57.57	52.60	23.91
S.E. _{est.} (Raw Score Points)		1.07	2.25	1.27	1.80	2.63	2.78	2.01	2.27	1.77	2.57	2.78	2.37
Approx. T-score change at 1 S.E.		3	4	3	4	7	5	5	4	6	4	4	5
F = Variance Ratio		1.45	1.17	1.39	1.05	1.42	1.00	1.10	1.00	1.03	1.24	1.25	1.14

* Rounded from four to two places after computations were completed.

that in two out of three administrations, the raw score earned on one form will be within ± 2.78 of the score that would be secured had the alternate form been given. This in sum is the meaning of the correlation coefficient (r'_{cb}) of .85 derived for the Sc scale.

The next to last row of Table 3 indicates the approximate change in T -score points for each shift of one standard error. Using the Sc scale again as an example, a change in raw score points of this magnitude results in a change of approximately four-tenths of a standard deviation. Such a change could be significant.

Profiles of means shown in Figures 1 and 2 indicate small T -score differences between card and booklet form. They could be considered normal profiles for a college group.

Summary

Correlation coefficients, means, and the standard deviations were computed for 68 male and 32 female college undergraduate and graduate students to determine the equivalence of the card and booklet forms of the MMPI. With the exception of L, D, and Pa, these coefficients range between .72 and .91. This is as high as the majority of the reliability coefficients reported for the card form alone.

Mean scores indicated in Table 2 suggest that a slightly more elevated profile would be secured on a college population by use of the booklet form. Fewer items are left unanswered on the booklet form.

Inclusion of the standard error of estimate in raw score points and the approximate T -score change at one standard error in Table 3 permit

one to judge whether he wishes to use these inventories as alternate forms. If these coefficients are interpreted as reliability indices, the standard error of estimate is as shown in Table 3. If the negatively skewed distribution of errors is recalled, that is, the tendency to score slightly higher on the booklet form, one is forced to adjust individual diagnosis upward from card to booklet form or resign himself to an inequivalence of forms.

Received October 21, 1949.

References

1. Wiener, D. N. Differences between the individual and group forms of the Minnesota Multiphasic Personality Inventory. *J. consult. Psychol.*, 1947, 11, 104-106.
2. Buros, O. K. *The third mental measurements year-book*. New Brunswick, Rutgers University Press, 1949. Comments of Julian B. Rotter on the Minnesota Multiphasic Personality Inventory on p. 60.
3. Hathaway, S. R., and J. C. McKinley. A Multiphasic Personality Schedule (Minnesota): I. Construction of the schedule. *J. Psychol.*, 1940, 10, 249-254; II. A differential study of hypochondriasis. *J. Psychol.*, 1940, 10, 255-268; III. The measurement of symptomatic depression. *J. Psychol.*, 1942, 14, 73-84; IV. Psychasthenia. *J. appl. Psychol.*, 1942, 26, 614-624; and V. Hysteria, hypomania, and psychopathic deviate. *J. appl. Psychol.*, 1944, 28, 153-174.
4. Holzberg, J. D., and Allessi, S. Reliability of shortened MMPI. *J. consult. Psychol.*, 1949, 13, 288-292.
5. Capwell, D. F. Personality patterns of adolescent girls. I. Girls who show improvement in I.Q. *J. appl. Psychol.*, 1945, 29, 212-228; and II. Delinquents and non-delinquents. *J. appl. Psychol.*, 1945, 29, 289-297.

A Factor Analysis of MMPI and Aptitude Test Data *

Lt. Comdr. Ellsworth B. Cook (MSC) U.S.N.

Tufts Medical School, Boston, Mass.

and

Robert J. Wherry

Department of Psychology, Ohio State University

This paper discusses the results of the administration of psychometric and psychomotor tests to a group of 120 naval enlisted submarine candidates. The study comprised one phase of an investigation of the possible value of a wide variety of measures for the selection of submarine personnel (1).

Procedure

Subjects were randomly selected (1) and two groups of six subjects each were tested weekly. The tests employed were:

I. The Minnesota Multiphasic Personality Inventory: Hereinafter referred to as the MMPI, this test is designed to provide scores on all the more important phases of personality (2, 3, 4, 5, 6) and has been used extensively for the overall differentiation of normals from abnormals or persons predisposed to abnormal developments (7, 8, 9, 10, 11, 12). The short group form was used.

II. Two-Hand Coordination Test: This is a motor pursuit task which has been employed frequently in the selection of military personnel (14, 15, 16, 17, 18, 19). The essential psychological principle involves the carrying out of two coordinated movements simultaneously so that there is a conflict of attention. The subject is rated on his ability to manipulate hand cranks in such a way as to keep a small button in continuous contact with an irregularly moving disc. An electrically operated stop clock measures the total amount of time during which actual contact is maintained. Two 1-minute trials were used.

* The study reported herein was conducted at the U.S.N. Medical Research Laboratory, U.S.N. Submarine Base, New London, Connecticut, under BuM&S Research Project NM-003-017.

Opinions expressed are those of the authors and are not to be construed as necessarily reflecting the views or the endorsement of the Navy Department.

III. Basic Battery of Written Tests: This battery consisted of a test of arithmetical reasoning (fractions, percentages, proportions, etc), mechanical and electrical knowledge (picture identification tests), mechanical aptitude (simple principles of physics—levers, pulleys, braces, etc.), and the General Classification Test (verbal abilities). These are standard Navy tests for enlisted men (20, 21). In order to qualify for Submarine School, enlisted candidates must have a combined score of 100 on the GCT and arithmetic tests (22).

IV. Navy Enlisted Personal Inventory: This consisted of form 2 of the Personal Inventory (23), a group test which presents a standardized psychiatric interview in pencil and paper form. The forced-choice type items which comprise the inventory are based on case history dissimilarities between psychiatrically undesirable and normal military personnel (24, 25, 26). Inasmuch as individual interviews must necessarily be brief during large scale selection programs, the P.I. serves as a rough screening device to guide the psychiatrist in orientating his interview. Scores on the two sections (personal history and medical history) were treated as separate variables.

V. Tank Performance. Subjects were rated on a five-point scale by a submarine medical officer for their overall performance while undergoing routine training procedures at the Escape Training Tank. This device, employed to acquaint personnel with the method of escaping from a submerged submarine, is a tower containing a column of fresh water 25' in diameter and 100' deep. A training bell permits an ascent from any desired depth, and hatches or locks are located at various points in the tower. Subjects made two underwater ascents from each of the 12', 18' and 50' depths. They were rated on such items as evidence of

Table 1
Means and Standard Deviations of Variables Selected for Analysis

Var. No.	Variable Description	Mean	S.D.
01	Lie Score	54.10	5.91
02	F (Validity) Score	52.88	4.68
03	Hs (Hypochondriasis) Score	46.58	6.83
04	D (Depression) Score	48.26	7.67
05	Hy (Hysteria) Score	52.73	6.89
06	Pd (Psychopathic Deviate) Score	52.40	8.53
07	Mf (Interest) Score	51.27	8.91
08	Pa (Paranoia) Score	47.84	7.22
09	Pt (Psychasthenia) Score	44.78	6.46
10	Sc (Schizophrenia) Score	46.77	6.29
11	Ma (Hypomania) Score	58.55	8.42
12	Two Hand Coordination (C Score)	11.75	1.88
13	General Classification Test	58.88	6.30
14	Arithmetical Reasoning	56.92	9.16
15	Mechanical Aptitude	57.77	6.77
16	Mechanical Knowledge	54.55	8.30
17	Electrical Knowledge	54.92	8.23
18	Tank Performance Grade	2.89	0.43
19	Personal History	1.18	1.46
20	Medical History	0.08	0.30

apprehension, quickness of response to instructions, errors of position on the line, "freezing" on the line, fighting to get out of the water too quickly, and so forth.

Statistical Analysis and Results

Data on 111 of the 120 subjects were utilized for statistical analysis. Three men were dropped because application of the standard criteria indicated that their MMPI scores were invalid. Six others were excluded because records on them were incomplete. The 20 variables selected for analysis are listed in Table 1, together with their means and standard deviations.

A comparison of mean scores for the 9 personality scales with those obtained by roughly similar groups (27, 28) indicated that performance on the MMPI was typical of a young male adult population.

The mean C score of 11.8 on two-hand coordination placed the group almost one standard deviation above the mean of the sample on which this test was validated for naval use (29), while the standard deviation of 1.9 was comparable to that of the standardizing group (29).

Scores were better than average on all the items of the Navy basic battery of written tests. This was to be expected since standards for the submarine service are higher than for the Navy generally (22). Mean scores closely approximated those of several hundred experienced submariners who were reassigned to New London in the summer of 1945 (30).

Mean scores were well below the established cut-offs for both sections of the Navy Personal Inventory.

A tabulation of subject score range on each personality scale of the MMPI (Table 2), a graphical presentation of the MMPI profiles of individuals who scored 70 or above on two or more scales compared with the mean of the whole group (Figure 1), and a comparison of subject performance on two-hand coordination with several other submarine populations (Table 3) are available on request.¹ Space does not permit their inclusion here.

¹ To reduce printing costs, Tables 2 and 3 and Figure 1 have been deposited with the American Documentation Institute. Order Document 2828 from American Documentation Institute, 1719 N Street, N.W., Washington 6, D. C., remitting \$.50 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$.50 for photocopies (6 X 8 inches) readable without optical aid.

The intercorrelations of the 20 variables are presented in Table 4. A modified (31) Thurstone Group Centroid (32) factor analysis yielded six independent factors to explain the intercorrelations obtained. The residuals arising when one attempts to explain the intercorrelations on the basis of the factor loadings are also included in Table 4. The factor loadings for the variables are given in Table 5. A factor loading represents the correlation between a given measurement and one of the factors isolated. It may be positive or negative depending on the nature of the relationship with the particular variable involved. The factor loading squared gives the percentage of score variance of a given measurement which may be explained or predicted by the factor in question. A loading of .20 or higher is regarded as significant. The reader is reminded that the labelling of factors is a matter of interpretative judgment rather than a problem in

statistics, and that he is free to consider and suggest alternate designations.

Factor A has high positive loadings on the validity (.64), hypochondriasis (.79), psychasthenia (.72) and schizophrenia (.93) scales of the MMPI, and lower but still significant loadings on the depression (.28), hysteria (.28), psychopathic deviate (.33), masculinity-femininity interest (.33), paranoia (.25), and hypomania (.41) scales of the MMPI, as well as on the personal (.21) and medical (.38) history sections of the Navy Personal Inventory. In general, then, it has significant projections on all items which measure neurotic tendencies, and is labelled *tendency to personality maladjustment*. The word "tendency" is employed to emphasize that the group was a normal one. Factor A appears comparable to the general factor "maladjusted tendencies" isolated by Cottle in his study of the MMPI and the Bell Adjustment Inventory (33).

Table 5
Final Factor Loadings

Variable Number and Description	Tendency to Personality Maladjustment Factor A	Numerical-Verbal Intelligence Factor B	Tendency to Over-Activity Factor C	Tendency to Paranoia Factor D	Mechanical Coordination Factor E	Tendency to Femininity of Interest Pattern Factor F	Communality h ₂
01 MMPI Lie Scale	-.06	-.26	-.34	.67	.17	-.09	.67
02 MMPI Validity Scale	.64	-.03	.02	-.11	.31	-.08	.53
03 MMPI Hypochondriasis Scale	.79	-.19	-.23	.03	.08	.00	.72
04 MMPI Depression Scale	.28	-.05	-.48	.12	.08	.01	.33
05 MMPI Hysteria Scale	.28	-.08	-.43	.34	.09	-.35	.52
06 MMPI Psychopathic Deviate Scale	.33	.16	-.38	-.04	-.01	.07	.29
07 MMPI Interest Scale	.33	.16	.01	.38	-.15	.39	.45
08 MMPI Paranoia Scale	.25	-.01	-.17	.48	-.14	.00	.34
09 MMPI Psychasthenia Scale	.72	.03	.10	-.09	-.01	.47	.76
10 MMPI Schizophrenia Scale	.93	.19	.14	-.08	.13	.18	.98
11 MMPI Hypomania Scale	.41	.09	.56	-.04	.09	-.01	.50
12 Two-Hand Coordination Test	-.06	.00	.10	-.13	.36	-.08	.17
13 General Classification Test	-.02	.76	.06	-.17	.25	.01	.67
14 Arithmetical Reasoning	-.04	.71	.05	.13	.08	.01	.53
15 Mechanical Aptitude	.12	.24	.34	.14	.43	-.41	.56
16 Mechanical Knowledge	-.02	.00	.38	-.01	.57	.12	.48
17 Electrical Knowledge	-.01	.07	.01	-.04	.68	-.05	.47
18 Tank Performance Grade	.09	-.06	-.19	-.09	-.01	-.02	.06
19 Personal History	.21	.10	.07	.16	.00	.40	.24
20 Medical History	.38	-.13	.12	.08	.05	.09	.19

Factor B has high positive loadings on the GCT (.76) and arithmetic (.71) tests and a lower positive loading (.24) for mechanical aptitude. This factor appears indicative of the ability to follow directions, and akin to the trait measured by traditional intelligence tests. Accordingly it is designated *numerical-verbal intelligence*. The factor has a significant negative loading (−.26) on the lie index of the MMPI, implying that persons who do well in intelligence tests tend to refrain from falsifying answers on personality tests.

Factor C has its highest loading on the hypomania scale of the MMPI (.56) and significant positive loadings on mechanical aptitude (.34) and mechanical knowledge (.38) as well. This is a logical pattern in that overactive individuals often find outlet in mechanical pursuits. The factor is called *tendency to over-activity*.

Over-active persons possess a considerable degree of emotionality (as evidence the negative loading of −.38 on the psychopathic deviate scale); this emotionality is shallow but varied.

The factor has significant negative loadings on the "neurotic triad"—the hyperchondriasis (−.23), depression (−.48) and hysteria (−.43) scales of the MMPI—indicating that individuals high on this factor tend to lack self-consciousness and self-criticism and have a direct acceptance of the environment. This suggestion of a "recklessness pattern" among men interested in submarine duty is somewhat similar to the finding of an Air Forces study of the traits of fighter pilots (34).

It is interesting to note that factor C has nearly zero loadings on the two-hand coordination test, although one would normally expect a correspondence between mechanical aptitude and two-hand coordination. The over-productivity in thought and action is evidently sufficient here to cause an attempt to think ahead, to "beat" the gadget by anticipating its movements, and, actually, to result in poor coordination performance.

The negative loading of −.19 on tank performance grade shown for this factor is worthy of mention, even though the loading is just under the established criterion (.20) of significance. Tank performance rating penalizes a man who "rushes" the line in an attempt to complete an ascent too quickly. Here again,

the element of impatience and impulsiveness appears. The finding is suggestive in view of a wartime service report (35) issued after a submarine crew had been subjected to long submergence and heavy depth charging. In the colorful language of that report: "... when the long dive was over . . . the people who lasted out were those of a more phlegmatic disposition who didn't bother much when things were running smoothly. The worriers and hurriers had all crapped out, leaving the plodders to bring home the ship" (35).

Factor D is labelled *tendency to paranoia* from the loading of .48 on the paranoia scale of the MMPI. The high loading on the lie index of the MMPI is logical in that individuals tending toward that trait approach personality tests suspiciously, and are prepared to admit nothing which might show them in an unfavorable light. The loading of .38 on the interest scale suggests that the individuals high on factor D were the more effeminate members of the group. There is a negative not quite significant loading on GCT, suggesting that those who falsify on the lie questions of the MMPI do poorly on GCT. Thus, factors B and D give corroborative support to one another. There may well be an index of stupidity present here also, with the less intelligent men falling more easily into the trap presented by the lie questions.

Factor E has its highest loadings on electrical knowledge (.68), mechanical knowledge (.57), mechanical aptitude (.43), and two-hand coordination (.36), and accordingly it is designated as *mechanical coordination*. The factor has a significant positive loading also on the validity scale of the MMPI (.31) indicating that persons high in mechanical coordination were meticulous in answering the questions of the personality test. The negative loading on the interest scale (−.15), while not quite significant, implies that the more masculine members of the group were more proficient mechanically. Factor E indicates also that the expected correspondence between mechanical ability and two-hand coordination is present when loadings on neurotic items are negligible, as is the case here.

Factor F has positive significant loadings on the masculinity-femininity interest scale (.39), the psychasthenia scale (.47) and the personal

history section of the P.I. (.40). The significant negative loading on mechanical aptitude (-.41) is taken to indicate that a man leaning toward the feminine side of the interest scale is likely to get a lower score in mechanical tasks than will a person whom this scale measures as more positively masculine in interests. This supports the Terman-Miles view that there is a pronounced relationship between masculinity and mechanical pursuits at every educational level (36), and Strong's definition of masculinity scores as an interest in things or objects rather than in persons or personalities (37). The most likely designation for factor F appears to be *tendency to femininity of interest pattern*. The high loading on psychasthenia shown for this factor suggests that the more effeminate man tends toward compulsive behavior; this is consistent with the MMPI test development where this is regarded as more a feminine than a masculine trait (13).

One purpose of each area study of the type reported herein is to select single or composite measures, of the truly basic factors isolated, for relational analysis to significant components found in other areas (1). All six factors isolated in this area investigation will be represented in the final matrix.

Summary

1. Men who do well on GCT and arithmetic tests tend to refrain from falsifying on the lie questions interspersed throughout the MMPI.
2. When neurotic elements are absent, there is the expected correspondence between mechanical ability and two-hand coordination. However, when such elements are present in the personality, they tend to make an individual rigid and confused with two-hand coordination.
3. A man whom the MMPI estimates as tending toward femininity of interest pattern is likely to do less well on tests of mechanical ability than will one whom the Mf interest scale measures as more positively masculine.
4. A trait suggestive of over-activity and recklessness was found in the personality pattern of some subjects. While alertness and daring are pre-requisites for successful submarine action, there is some evidence that over-active individuals may find it difficult to

tolerate prolonged submergence and confinement. Presumably the "hurrier" and the "plodder" both have place in the complete scheme of underseas operation, with its long periods of monotony interspersed with moments of intense activity.

5. The evidence of a relationship between performance on intelligence and aptitude tests with personality traits as measured by the MMPI is considered worthy of note inasmuch as the minor personality accentuations found in this sample were within the generally acceptable ranges.

Received September 26, 1949.

References

1. Cook, E. B., and Wherry, R. J. A study of the interrelationships of psychological and physiological measures on submarine enlisted candidates: I. History, experimental design and statistical treatment of data. Report No. 1 BuM&S Research Project NM-003-017, U. S. Naval Medical Research Laboratory, U. S. Naval Submarine Base, New London, Conn., 9 March 1949.
2. Hathaway, S. R., and McKinley, J. C. A multiphasic personality schedule: I. Construction of the schedule. *J. Psychol.*, 1940, 10, 249-254.
3. McKinley, J. C., and Hathaway, S. R. A multiphasic personality schedule: II. A differential study of hypochondriasis. *J. Psychol.*, 1942, 10, 255-268.
4. Hathaway, S. R., and McKinley, J. C. A multiphasic personality schedule: III. The measurement of symptomatic depression. *J. Psychol.*, 1942, 14, 73-84.
5. McKinley, J. C., and Hathaway, S. R. A multiphasic personality schedule: IV. Psychasthenia. *J. appl. Psychol.*, 1942, 26, 614-624.
6. McKinley, J. C., and Hathaway, S. R. A multiphasic personality schedule: V. Hysteria, hypomania and psychopathic deviate. *J. appl. Psychol.*, 1944, 28, 153-174.
7. Morris, W. W. A preliminary evaluation of the Minnesota Multiphasic Personality Inventory. *J. clin. Psychol.*, 1947, 3, 370-374.
8. Hunt, H. F., Carp, A., et al. A study of the differential diagnosis efficiency of the Minnesota Multiphasic Personality Inventory. *J. consult. Psychol.*, 1948, 12, 331-336.
9. Schiele, B. C., Baker, A. B., and Hathaway, S. R. The Minnesota Multiphasic Personality Inventory. n.d. Departments of Neuropsychiatry and of Psychology, University of Minnesota Medical School.
10. Meehl, P. E. Profile analysis of the Minnesota Multiphasic Personality Inventory in differential diagnosis. *J. appl. Psychol.*, 1946, 30, 517-524.

11. Clark, J. H. Application of the MMPI in differentiating A.W.O.L. recidivists from non-recidivists. *J. Psychol.*, 1948, 26, 229-234.
12. Abramson, H. A. The Minnesota personality test in relation to selection of specialized military personnel. *Psychosom. Med.*, 1945, 7, 178-184.
13. Hathaway, S. R., and McKinley, J. C. Manual of the Minnesota Multiphasic Personality Inventory, Revised Edition. New York: The Psychological Corporation, 1943.
14. McFarland, R. A., and Channell, R. C. A two-hand coordination apparatus for appraising aptitude for flying. Division of Research, C.A.A., Washington, D. C., March 1942.
15. Anon. The two-hand coordination test performance of submarine men. Brown University, Providence, R. I. Report No. 5, project 44, Section D-4, NDRC, September 1942.
16. Graham, C. H., Riggs, L. A., Bartlett, N. R., et al. A report of research on selection tests at the U. S. Submarine Base, New London, Conn. Brown University, Providence, R. I. OSRD report No. 1770, project 44, Division 7, June 1943.
17. McFarland, R. A., and Channell, R. C. A revised two-hand coordination test. Airman Development Division, C.A.A., Washington, D. C. Report No. 36, October 1944.
18. McFarland, R. A., and Franzen, R. The Pensacola study of naval aviators. Division of Research, C.A.A., Washington, D. C. Report No. 38, November 1944.
19. NRC Comm. on Selection and Training of Aircraft Pilots. Report on the Boston-Midwest project. Division of Research, C.A.A., Washington, D. C., Report No. 52, November 1945.
20. U. S. Navy. Arithmetical reasoning test and mechanical aptitude test. Bureau of Naval Personnel, Training Standards Section, Standards and Curriculum Division, Test and Research Unit. NavPers 16992, December 1944 (R).
21. U. S. Navy. Electrical knowledge test, mechanical knowledge test, general classification test. Bureau of Naval Personnel, Training Standards Section, Standards and Curriculum Division, Test and Research Unit. NavPers. 16994, December 1944 (R).
22. Willmon, T. L. Outline and discussion of methods for selection of submarine reserve personnel. U. S. Naval Medical Research Laboratory, U. S. Naval Submarine Base, New London, Conn. 16 February 1948.
23. U. S. Navy. Navy enlisted personal inventory, form 2, NavPers. 16845, IBM Form I.T.S. 1100 A 1165 (R).
24. Shipley, W. C., Gray, F., and Newbert, N. Standardization and validation of the personal inventory: psychiatric criterion. OSRD report No. 1606, Brown University, Providence, R. I. June 1943.
25. Shipley, W. C., and Graham, C. H. Final report in summary of research on the personal inventory and other tests. Applied Psychology Panel, NDRC, Report No. 10, project N-113, August 1944.
26. Kogan, L. S., Wantman, M. J., and Dunlap, J. W. Analysis of the personal history inventory. Division of Research, C.A.A., Washington, D. C. Report No. 42, February 1945.
27. Wiener, D. N. Differences between the individual and group forms of the MMPI. *J. consult. Psychol.*, 1947, 11, 104-106.
28. Clark, J. H. Some MMPI correlates of color responses in the group Rorschach. *J. consult. Psychol.*, 1948, 12, 384-386.
29. Bartlett, N. R. Review of research and development in examination for submarine training 1942-1945. Report No. 2, BuM&S Research Project NM-003-036, U. S. Naval Medical Research Laboratory, U. S. Naval Submarine Base, New London, Conn. (In preparation.)
30. Bartlett, N. R. Report on correlations of tests with grades in submarine school. Report No. 2, BuM&S Research Project X-243 (sub. 47), U. S. Naval Medical Research Laboratory, U. S. Naval Submarine Base, New London, Conn. 13 February 1945.
31. Wherry, R. H., Brogden, H. E., and Gaylord, R. H. Wherry-Brogden, Gaylord method of factor analysis. Personnel Research Section, Adjutant General's Office, Department of the Army, Washington, D. C. (unpublished).
32. Thurstone, L. L. *Multiple factor analysis*. Chicago: University of Chicago Press, 1947.
33. Cottle, W. C. A factorial study of selected instruments for measuring personality and interest. Guidance Bureau, University of Kansas, n.d.
34. U. S. A. A. F. Psychological research on operational training in the continental air forces: AAF Aviation Psychology Program, Report No. 16, Washington, D. C.: U. S. Government Printing Office, 1947.
35. U. S. Navy. Depth charging of the U.S.S. Puffer. Section 71 T of report, Enemy anti-submarine measures. n.d.
36. Terman, L. M., and Miles, C. C. *Sex and personality*. New York: McGraw-Hill Co., 1936.
37. Strong, E. K., Jr. *Vocational interests of men and women*. Palo Alto: Stanford University Press, 1943.

A Combined Oral Reading and Psychogalvanic Response Technique for Investigating Certain Reading Abilities of College Students

Homer L. J. Carter

Western Michigan College

The purpose of this study is to describe and evaluate a combined oral reading and psychogalvanic response technique for investigating certain reading abilities of college students. It is postulated that the procedure may be of value not only in determining such factors as reading rate, comprehension, and errors, but as a means of discovering how much the reading situation affects the individual. In therapy this is important for as Maier (4) has shown frustration state, not behavior symptoms, must be treated. In order to evaluate this procedure, the materials and apparatus have been described, superior and inferior readers have been studied, resulting group data have been compared statistically, and inferences have been set forth tentatively.

Materials and Apparatus

(1) *Gray Oral Reading Paragraphs Test* (3). Six of the even numbered paragraphs were taken from the Gray Oral Reading Paragraphs Test. These paragraphs were typed on 3"X5" cards, and on the back of each card five questions were typed. It is assumed that the paragraphs given in the order 2, 4, 6, 8, 10, 12 constitute a scale of increasing difficulty but with steps of greater magnitude than that of the original scale from which the paragraphs were chosen. This modification of the Gray Oral Reading Paragraphs Test provides a record of reading rate, a comprehension score, and such reading errors as words aided, mispronounced, omitted, substituted, inserted and repeated.

(2) *Apparatus*. The apparatus used throughout this experiment was a "Maico Psychometer" (1). Change in palmar skin resistance (Delta R) can be measured in units ranging from zero to 100 and stated in ohms by merely multiplying the indicated unit on the scale by 200. In this experiment, only change in palmar skin resistance in response to given

stimuli is considered and it has been assumed that the extent of the deflection of the galvanometer is roughly proportionate to the intensity of the emotion or degree of frustration.

Procedure

Selection of Students. Twenty superior readers were selected from students scoring above the 75th percentile on Test III (reading) of the Ohio State Psychological Examination and twenty inferior readers were chosen from those scoring below the 25th percentile. Such factors as age, sex, and academic training were considered in making up the groups of superior and inferior readers. No attempt was made to control the factors of scholastic aptitude or general intelligence. First, second, third, and fourth year college students constituted both groups although the number of freshmen was equal to that of all upperclassmen.

Application of Technique. As the apparatus was being applied, the examinee was given paragraph 2 and the following directions, "Read the paragraph on the card aloud. Avoid all reading errors. After you have finished, you will be asked questions concerning the material read." The response to each paragraph was recorded in terms of number of errors, time required for reading, change in palmar skin resistance and comprehension score. A record of these data in the case of a freshman with a percentile of 8 on the reading section of the Ohio State Psychological Examination is shown in Table 1.

Results

Data resulting from the administration of a combined oral reading and psychogalvanic response technique in the study of superior and inferior readers at the college level have been summarized as shown in Table 2. In deter-

Table 1
Data Resulting from Application of Technique
in Individual Case

Para- graph	Errors	Time in Sec.	Delta R	Compre- hension (Weighted Score)
4	0	17	10	5
6	1	19	16	6
8	0	21	24	6
10	4	30	32	0
12	5	28	32	0

mining scores in comprehension, weighted values of 1 to 5 were assigned to questions on paragraphs 4 through 12, respectively. Because paragraph 2 was used in preparation for the examination, data resulting from its use were not included. The mean and sigma were determined for each distribution and the standard error of the difference of the means of small samples was found (2). In determining whether or not the differences between groups were significant, t was calculated for each difference. Differences between the means in the tabulation of the total number of errors, average time and comprehension are significant at the 1 per cent level. The average change in palmar skin resistance as shown by superior and inferior readers as they read paragraphs 4, 6, 8, 10, and 12 is not statistically significant ($t=.66$). However, on the more difficult paragraphs 8, 10, and 12 the difference in the means is statistically significant ($t=1.80$) at a point between the 5 and 10 per cent levels.

An analysis of data resulting from this study shows three trends which may be significant. Eight good readers and thirteen poor

readers show an increase in frustration as number of errors increased and a loss in comprehension occurred. This suggests that more poor readers are frustrated by these reading disabilities than good readers. It is also apparent that 7 good readers and 6 poor readers show a decrease in frustration as errors increased and a loss in comprehension occurred. Consequently it may be that some individuals in both groups are not emotionally affected by their errors and inability to comprehend. Furthermore, 1 poor reader and 2 good readers who show increased frustration demonstrate decreases in both errors and in comprehension. This may indicate that in their cases emotional tension is due only to inability to understand what is read.

Summary

1. As generally expected such factors as number of errors, rate of reading and comprehension scores differentiate significantly superior and inferior readers.

2. Average change in palmar skin resistance cannot be expected to differentiate superior and inferior readers except as the material becomes comparatively more difficult.

3. Nevertheless, this study is significant because it suggests the importance of applying measures of frustration simultaneously with measurement of reading achievement. Consideration of frustration in human behavior is in keeping with the contributions of Maier in his studies of animal behavior.

4. The technique described in this study provides objective data as to how much the accumulative effect of certain reading errors and inabilities affects the reader. This information may be of value not only in the diag-

Table 2
Data Resulting from Administration of Technique to Good and Poor Readers

	Good Readers		Poor Readers		S.E.D	t
	Mean	Sigma	Mean	Sigma		
Total Number of Errors	4	3.29	13	5.85	1.54	5.84
Average Time in Seconds	18.75	2.30	23.7	4.73	1.20	4.12
Comprehension Score	63.75	14.08	31.5	11.28	4.14	7.79
Average Change in Palmar Skin Resistance for Paragraphs 8, 10, 12	29	17.18	39.5	18.83	5.85	1.80

nosis of reading disability but in its correction as well. As a result of this information, therapy can be directed toward the reduction of frustration, for example by providing easier material, and not merely toward the eradication of behavior symptoms such as rate, errors and comprehension.

Received May 1, 1950.

Early publication.

References

1. Carter, H. L. J. A combined projective and psychogalvanic response technique for investigating certain affective processes. *J. consult. Psychol.*, 1947, 11, 270-275.
2. Garrett, H. E. *Statistics in psychology and education*. New York: Longmans, Green, 1947.
3. Gray, William S. *Standardized oral reading paragraphs*. Bloomington, Illinois: Public School Publishing Company.
4. Maier, Norman R. F. *Frustration*. New York: McGraw-Hill Book Company, Inc., 1949.

Geographical Sampling in Testing the Appeal of Radio Broadcasts

John Gray Peatman

City College of New York

and

Tore Hallonquist

Columbia Broadcasting System

For more than a decade hundreds of national and local radio programs broadcast over the outlets of the Columbia Broadcasting System have been studied with the Program Analyzer method for the purpose of determining their strong and weak points and improving them in the light of listeners' reactions and comments. Until 1947 all of these tests were conducted in the New York studios and consequently the question arose as to whether audiences in other parts of the country would give reactions similar to those obtained from New York listeners. To get at least a preliminary answer to this question, CBS employed the Program Analyzer technique for eight weeks in Hollywood and two weeks in Boston during the latter part of 1947.

Thus, this article describes the results of a series of Program Analyzer tests with New York audiences, Los Angeles audiences, and Boston audiences. The New York and Los Angeles audiences were presented two *network* programs. One of these, a comedy-drama, originated in Hollywood and featured mainly West Coast talent. The other network program, a musical variety show, originated in New York and featured mainly East Coast talent. For the comparison of Boston and New York listeners an audience participation program *local* to Boston was used. One might expect considerable differences among these samples of the radio audience because of local familiarities and appeals. The Hollywood program, for example, might be expected to have greater appeal on the West Coast and the New York program to have greater appeal on the East Coast. Similarly, one might expect the local Boston program to have a much stronger appeal for Boston audiences than for New York audiences.

The Program Analyzer Technique

Before describing the results of the above comparisons and considering the question whether the Program Analyzer technique has general or only limited usefulness in testing the appeal of radio programs, we shall briefly review the Program Analyzer technique, originally developed in 1937 by Frank Stanton, now President of the Columbia Broadcasting System, and Paul Lazarsfeld, now Chairman of the Department of Sociology at Columbia University.

The Program Analyzer method brings a sample of listeners into direct contact with a radio or television program under scientifically controlled conditions and records the listeners' reactions to each successive second of the broadcast.

There are two versions of the Program Analyzer: "Little Annie" which has been continuously in operation at CBS since 1940 and "Big Annie" developed by CBS engineers and used at CBS since 1944.

"Little Annie" consists of a moving tape and a battery of 20 capillary pens. Ten of these pens draw continuous red lines which follow guide lines on the tape. The other ten pens draw lines in green. Each red and green pen is electrically connected with a red and green push-button in the Program Analyzer studio, so that when a button is pressed down the pen jogs off the guide line on the tape and stays off as long as the electrical circuit remains closed. Two "Little Annies" are used simultaneously and can record reactions of 20 listeners.

Regularly, broadcast spot announcements invite listeners to participate in a Program Analyzer test. People who respond to these announcements are classified on the basis of

sex, age, education, occupation, and availability at specific times. This information is punched on IBM cards, one card for each person. Each time a test is scheduled, the cards are run through the sorting machine to yield a sample of listeners that is controlled with respect to such factors as sex, age, and education—all of which have been found relevant for stratification sampling of radio listeners.

Listeners are invited in groups of from 10 to 20 people. They are seated in the Program Analyzer studio, offered cigarettes and put at ease. Each person is given a red push-button for the left hand and a green button for the right hand.

The group is usually given a pre-test questionnaire containing questions about home listening habits, program preferences, attitudes, and so forth. A recording played over the loud-speaker system next informs the listeners about the test procedure. They are asked to listen closely to the program, to press the green button and keep it pressed down when they think a program part is good—when they want to listen to it; to press the red button when they think a program part is poor—when they don't want to listen to it; to press neither button when they are indifferent to what they hear. The lights are dimmed and a recorded version of the program is then presented. One of the prime assets of the technique is that listeners, caught up in the momentum of the program, make spontaneous, non-reflective responses. The Program Analyzer, it may be said, provides a psychological situation favorable for eliciting a true response, even from listeners who normally would not be able to articulate their reactions and ideas. Because the procedure itself—the pressing of the buttons—is an extremely simple, spontaneous, and confidential procedure, the listener is almost irresistibly led on to take a stand in regard to the various aspects of the program. Once he has committed himself, oral articulation during subsequent interviewing is facilitated.

The interviewer in charge starts the recording apparatus in an adjoining room. As the program proceeds he notes down the positive and negative reactions of each listener, as indicated by the jogging pens. Thus, he knows at the conclusion of the show the

spontaneous reaction pattern of each subject. He goes back into the studio, the lights are turned on, the main questionnaire is distributed, and the attempt made to elicit the listeners' considered opinions about what they heard: their attitudes towards the program as a whole, what they liked and did not like about it, their reactions to specific elements, their opinions of the cast, etc.

Finally, there is a period of oral interviewing. First calling on a listener whose reactions to the program were unfavorable, next calling on a listener whose reactions indicated satisfaction with the show, the interviewer encourages each person to talk informally about the program. Thus, two opposing points of view are established right at the start so that no listener need feel that his opinion is in the minority. Each participant is asked to tell how he felt about each aspect of the program and, in turn, is asked to give his conscious reasons for his reactions recorded on the tape.

The questions and answers during the interview period are taken down in shorthand and transcribed.

Thus, the Program Analyzer technique ("Little Annie") yields three sets of data: (1) the second-by-second approval, disapproval, and indifference reactions of *each listener* as recorded on the Program Analyzer tape; (2) listener attitudes and opinions as expressed in writing in the test questionnaire; and (3) listener attitudes and opinions as expressed in the oral interview.

The Program Analyzer technique and the type of information obtained when "Big Annie" is used are the same as for "Little Annie," except that approval and disapproval reactions of the group as a whole are totalized instead of being differentiated for each listener. Furthermore, due to the size of the group present in the studio, it is not feasible to interview each subject at the end of the test session. Thus, the investigator relies mainly on the written questionnaire results for the verbalized opinions of the listeners.

The Hollywood Program

This national network comedy program originally developed by the CBS Network Program Department in Hollywood has main-

Table 1

Composition of the Two Samples for the Test of the Hollywood Program

	New York Sample N = 76	Los Angeles Sample N = 95
Sex:		
Male	42%	47%
Female	58	53
	100%	100%
Age:		
Under 26	26%	16%
26-40	41	34
Over 40	33	50
	100%	100%
Education:		
Grammar School	17%	31%
High School	61	40
College	22	29
	100%	100%

tained a very high listener appeal for a number of years as indicated by audience ratings.

A broadcast of the program was tested simultaneously with a New York sample of 76 listeners and a Los Angeles sample of 95 listeners. The Los Angeles sample was tested with "Big Annie" and the New York sample with "Little Annie" in six successive sessions with approximately 12 subjects at each session. The composition of each sample, with respect to sex, age, and education is shown in Table 1. Some of the subjects were regular listeners to this program, others had never heard it.

On the whole, the Los Angeles sample was more familiar with the program than the New York sample as shown in Table 2. The overall reaction of the listeners to the program, as measured by the Program Analyzer scores,¹ was identical for both samples, i.e., both tests yielded an average attitude score of 32 which is well above average for a comedy program.

But just as striking is the comparison of the minute by minute reactions of the two samples

¹ This average attitude score takes into account the proportion of listeners with (a) positive reactions; (b) indifferent or no reactions; and (c) negative reactions. See Peatman and Hallonquist, *The patterning of listener attitudes towards radio broadcasts*, Stanford University Press, 1945, page 33.

Table 2

Prior Listening Habits to the Hollywood Program

Listen to Program	New York Sample	Los Angeles Sample
Regularly (almost every broadcast)	23%	26%
Frequently (about every other week)	13	33
Occasionally (once in a while)	33	24
Never	31	17
	100%	100%

to the sequences of events on the broadcast. These reactions are portrayed in Figure 1. The trend of listener reactions of each sample is given by the heavy line graph on the two charts of Figure 1. This trend is measured by the Program Analyzer scores for successive units of the program. It will be observed that the trend lines for both samples are very similar, despite the fact that 31 per cent of the New York sample had not heard the program before. The Los Angeles sample responded more quickly in their approval of the program, a result probably attributable to the greater familiarity of the Los Angeles audience with the program and with its leading personality. The low points in each case are the commercials, a typical result in Program Analyzer tests of commercial programs. The over-all reaction of the listeners as obtained from the questionnaire administered at the end of each test session is practically the same for both samples, as indicated in Table 3. Conditional listeners are those who would listen to future broadcasts if the programs were "improved a bit." The main appeals of the program to both samples were very similar, as indicated in Table 4 which shows the relative appeal of the principal

Table 3

Per Cent Satisfied, Conditional, and Dissatisfied Listeners to the Hollywood Program

	New York Sample	Los Angeles Sample
Satisfied listeners	53%	55%
Conditional listeners	27	24
Dissatisfied listeners	20	21
	100%	100%

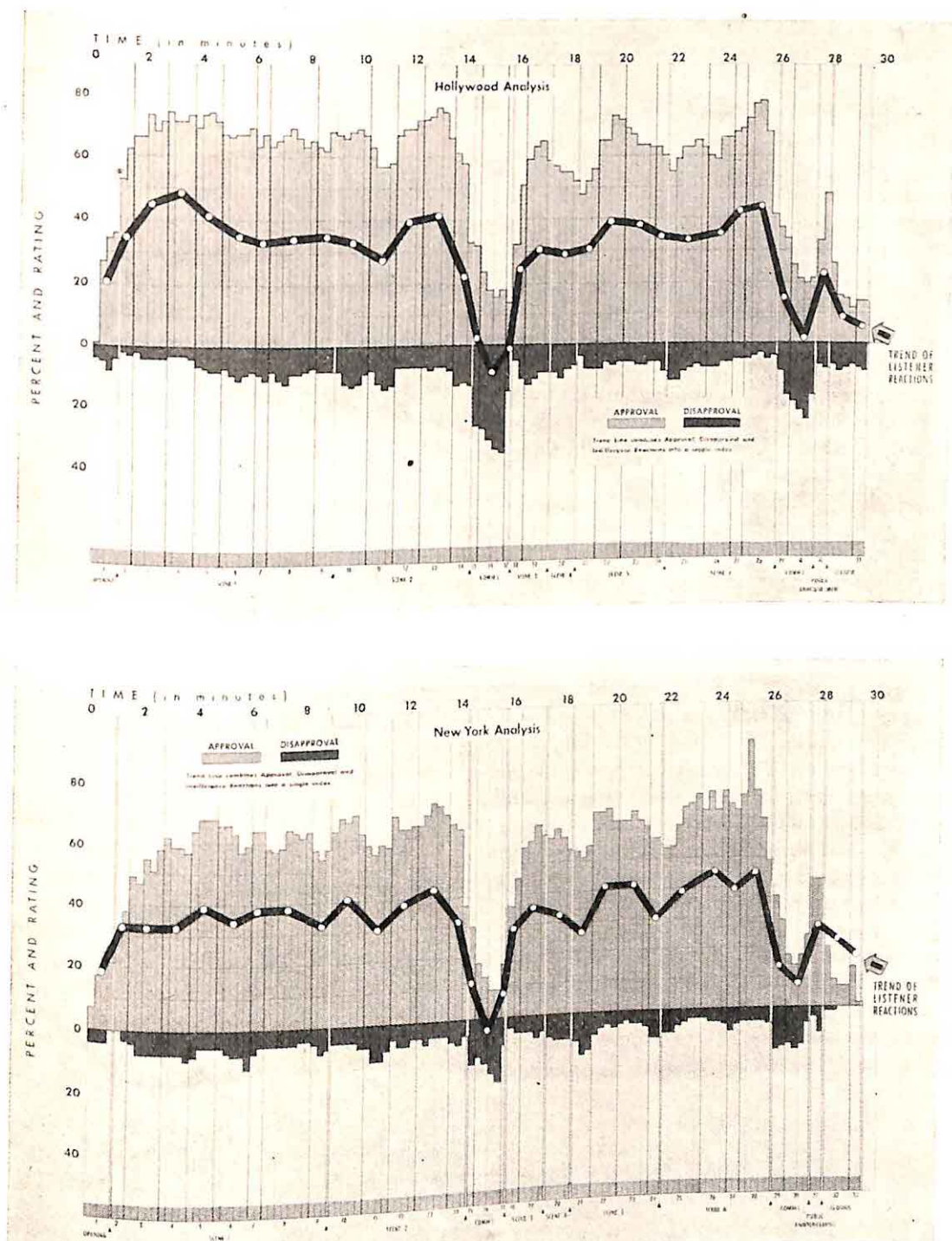


FIG. 1. Comparison of Hollywood and New York listeners: Hollywood comedy drama. Top—Profile of listener reactions, Hollywood listeners. Bottom—Profile of listener reactions, New York listeners.

Table 4

The Relative Appeal of Program Elements of the Hollywood Program

	New York Sample	Los Angeles Sample
The gags and jokes	49%	49%
Personalities and characters	37	38
Story and plot	9	9
Enjoyed none of these	5	4
	100%	100%

aspects of the broadcast. The story and plot had relatively the least appeal for both samples. It is to be emphasized that these figures describe the *relative* appeal of the program elements as obtained from the questionnaire which asked: "This is what I enjoyed most about the show":—The gags and jokes;—The story and plot;—The personalities and characters;—Enjoyed none of these. Each subject checked one answer to indicate which of the foregoing elements was most enjoyable.

The subjects were also asked to indicate whether or not they were anxious to hear the outcome of the plot. Their replies are described in Table 5.

Finally, the great similarity of the reactions of both samples to the broadcast is shown by Figure 2, which describes the comedy appeal of each of the characters on the program. Each percentage value is the per cent checking each character as "very funny." It will be observed that the ranks for each of the five characters are the same for both samples even though there are some differences in the percentage value for each.

From the foregoing figures and tables we see a remarkable similarity of both the over-all

Table 5

Interest in the Outcome of the Story of the Hollywood Program

	New York Sample	Los Angeles Sample
Anxious to hear outcome	72%	73%
Not anxious	28	27
	100%	100%

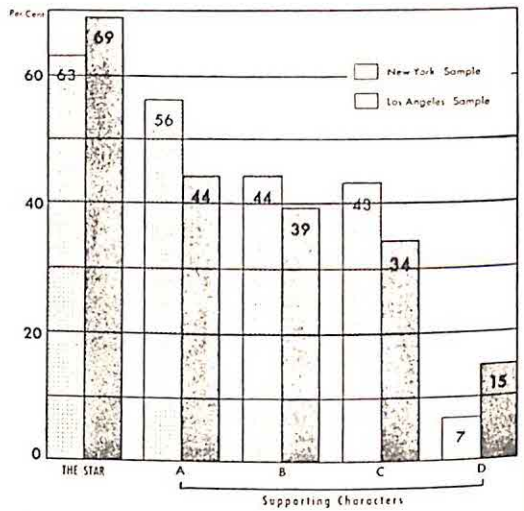


FIG. 2. Comedy appeal of program characters on Hollywood program (per cent checking each character as "Very Funny").

and detailed reactions of two different, geographically located samples hearing the same broadcast. Such identity of response certainly is not to be expected for all programs, particularly for programs that may have sectional or local flavor in their content. It is also to be noted that both samples were city samples and such similarity of response might not be ex-

Table 6

Composition of the Two Samples for the New York Variety Show

	New York Sample N = 71	Los Angeles Sample N = 94
Sex:		
Male	37%	49%
Female	63	51
	100%	100%
Age:		
Under 26	28%	22%
26-40	37	35
Over 40	35	43
	100%	100%
Education:		
Grammar School	13%	39%
High School	69	41
College	18	20
	100%	100%

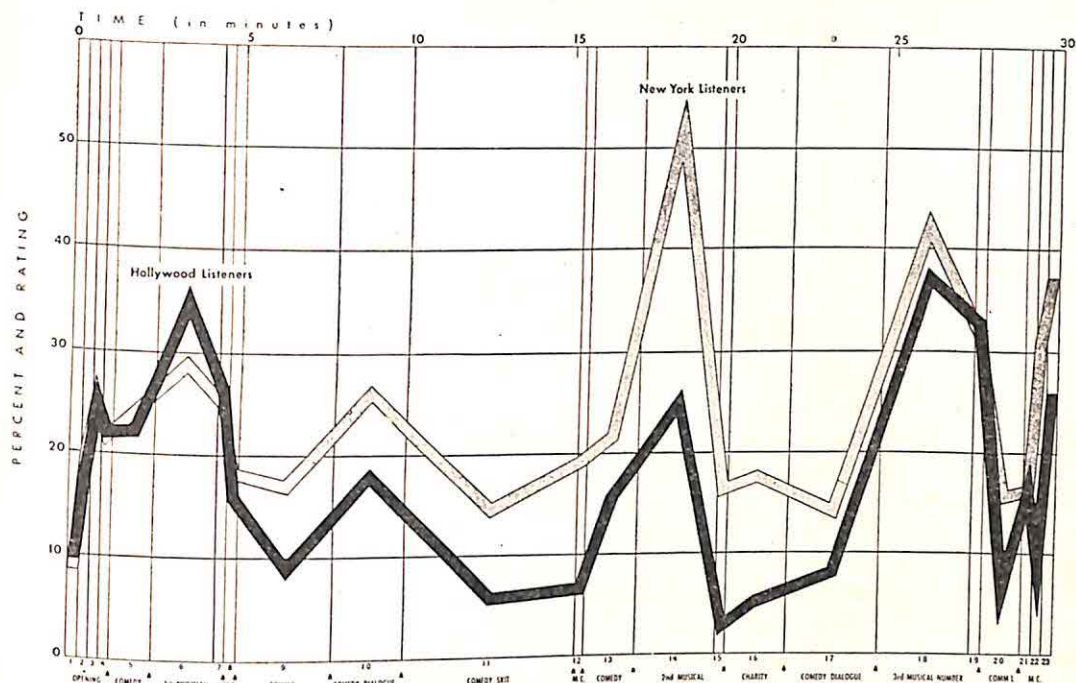


FIG. 3. Comparison of reaction trends for New York and Hollywood listeners, New York musical variety show.

pected in the comparison of rural or smalltown subjects with those of urban dwellers.

The New York Variety Show

In the second radio program to be considered here, comparisons again were made between New York listeners and Los Angeles listeners. This time, the program originated in New York City and its chief talent had obtained millions of East Coast fans long before the West Coast ever heard him. However, as in the case of the Hollywood program, the New York show has also developed a great deal of national

appeal over the CBS network. A particular broadcast was tested simultaneously in New York and Los Angeles. The two samples consisted of 71 listeners in the New York test and 94 listeners in the Los Angeles test. The composition of the two samples is given in Table 6.

As for previous familiarity with the program, we see in Table 7 that the situation is reversed from the Hollywood program samples: more of the Los Angeles sample had never heard the New York show before and a greater proportion of the New York sample listened regularly or at least "about every other week."

As in the preceding tests of the Hollywood program, "Big Annie" was used for the Los Angeles sample whereas "Little Annie" was used in a series of seven successive sessions with the New York sample. The over-all results of the tests, as measured by the Program Analyzer score, were not too dissimilar.

The reactions of the two samples to the successive program units of the broadcast are portrayed by the line graphs of Figure 3, each of which is based upon the average Program Analyzer scores for each program unit. The general trend of reactions for both samples

Table 7

Prior Listening Habits to the New York Variety Show

Listen to the Program:	New York Sample	Los Angeles Sample
Regularly (almost every broadcast)	35%	24%
Frequently (about every other week)	20	13
Occasionally (once in a while)	28	40
Never	17	23
	100%	100%

Table 8

Per Cent Satisfied, Conditional, and Dissatisfied
Listeners to the New York Variety Show

	New York Sample	Los Angeles Sample
Satisfied listeners	50%	39%
Conditional listeners	25	25
Dissatisfied listeners	25	36
	100%	100%

is essentially the same, but the *level* of the trend for the Los Angeles audience is somewhat lower than that for the New York sample. The high points and the low points in nearly all cases are found to be practically identical. In other words, the implications of the results for these two samples are very similar in that any recommendations for the improvement of the program, based on the Program Analyzer tests, would necessarily point to similar aspects of the broadcast.

Differences between the results of the two samples are brought out rather clearly in a comparison of the relative appeals for each of the program elements as shown in Table 9. The M.C. (Master of Ceremonies) has the most appeal for the New York sample whereas the musical numbers had relatively the greatest appeal for the Los Angeles sample.

Table 5, for the Hollywood comedy program, compared the two samples on the basis of their interest in the outcome of the plot. In the case of the New York musical variety show, there is, of course, no plot, but a comparison can be made of the listeners' judgment of the musical numbers and the comedy skit. Each

Table 9

The Relative Appeal of Program Elements—
New York Variety Show

	New York Sample	Los Angeles Sample
M.C.'s personality and humor	43%	28%
The musical numbers	40	52
Comedy dialogue	10	7
Comedy skit	6	6
Other appeals	1	7
	100%	100%

Table 10

Performances Liked Most and Liked Least

	New York Sample	Los Angeles Sample
Performances liked best:		
No. 1 Third musical number	42%	52%
No. 2 First musical number	23	18
No. 3 Comedy skit	21	17
No. 4 Second musical number	14	14
	100%	100%
Performances liked least:		
No. 1 Comedy skit	56%	58%
No. 2 First musical number	16	13
No. 3 Third musical number	14	14
No. 4 Second musical number	14	15
	100%	100%

subject was asked which performance he liked best and which performance he liked least. The results are brought together in Table 10.

It will be seen that, despite some differences in the trend of their over-all Program Analyzer reactions to the program and despite the difference in their reactions to the M.C., the two samples were in close agreement with respect to the individual performances. As is evident from the data of Table 10, they ranked the entertainers they liked best in the same order and they ranked the entertainers they liked least also in practically the same order.

The Boston Program

This *local* daytime audience-participation show was developed by CBS's owned and operated Boston station WEEI and had been heard for more than a year at the time the Program Analyzer tests were made. In the case of this comparison, therefore, the Boston sample was exposed to a program with which some were familiar and the New York sample was presented a program which none had heard previously. Both samples consisted of 71 listeners: "Little Annie" was used for both tests in seven successive sessions with approximately 10 subjects for each. The composition of the sample with respect to sex, age, and education was as shown in Table 11.

Inasmuch as the New York listeners had not

Table 11
Composition of the Two Samples—
for the Boston Program

	New York Sample N = 71	Boston Sample N = 71
Sex:		
Male	17%	3%
Female	83	97
	100%	100%
Age:		
Under 26	14%	12%
26-40	41	27
Over 40	45	61
	100%	100%
Education:		
Grammar School	17%	21%
High School	69	62
College	14	17
	100%	100%

heard the program before, a somewhat different type of question was asked for comparative purposes, namely, "How do you feel about quiz programs broadcast in the daytime?"

The Boston program featured two quizzes—a music quiz and a radio star quiz—as well as two contests in the nature of stunts—a doughnut-dunking contest and a cake-slicing contest. A special feature of the program was a "traveling mike," consisting of a search among the members of the studio audience for the woman with the most children. A total of 11 local contestants appeared on the program and were interviewed by the quizmaster prior to each contest. The prizes consisted of merchandise. A special program unit of the broadcast was the "Radio Mirror" plug—a reference to the fact that the program had received national recognition in the August issue of this magazine. The first half of this thirty-minute program was sponsored and included three commercials for a nationally advertised brand of bread. The second half was sustaining.

The over-all reaction of listeners to the program, as measured by their Program Analyzer scores, was as follows: Boston listeners 33 (a very good rating for this type of show), New York listeners 25 (about average). The pro-

gram thus had a stronger over-all appeal for Bostonians than for New Yorkers which is not unexpected in view of the local flavor of some of the content and subject matter characteristic of the show. This is true not only in terms of the over-all reaction of the two samples but also in terms of their reactions to the individual program units described in Figure 4.

The two trend lines are essentially parallel. The highspots and lowspots of the program are practically identical for both Boston and New York listeners. The musical quiz had the most appeal for both samples; the doughnut-dunking contest and the cake-slicing contest had the least appeal, aside from the commercials. The only difference of any consequence was in the relative appeal of the "plug" for Radio Mirror magazine which had some appeal for Boston listeners (local pride) but was of little interest to New York listeners.

The over-all reaction of listeners to the program, obtained from the questionnaire administered at the end of the test sessions, further established a difference between the reactions of the two samples, as indicated in Table 13.

A majority of the Boston sample was satisfied with the program whereas a third of the New York listeners were dissatisfied. Despite this fact, however, the main appeals of the program were similar for both samples. This has been seen in Figure 3 and is further confirmed in the questionnaire data summarized in Table 14.

Self-testing, which has been found to be a principal appeal in most quiz programs, was checked by nearly four-fifths of both groups as

Table 12
Opinions about Day-Time Quiz Programs

	New York Sample	Boston Sample
Such programs among my favorites	54%	43%
Like them as well as other daytime shows	31	39
Do not like them as well as other daytime shows or never listen to them	15	18
	100%	100%

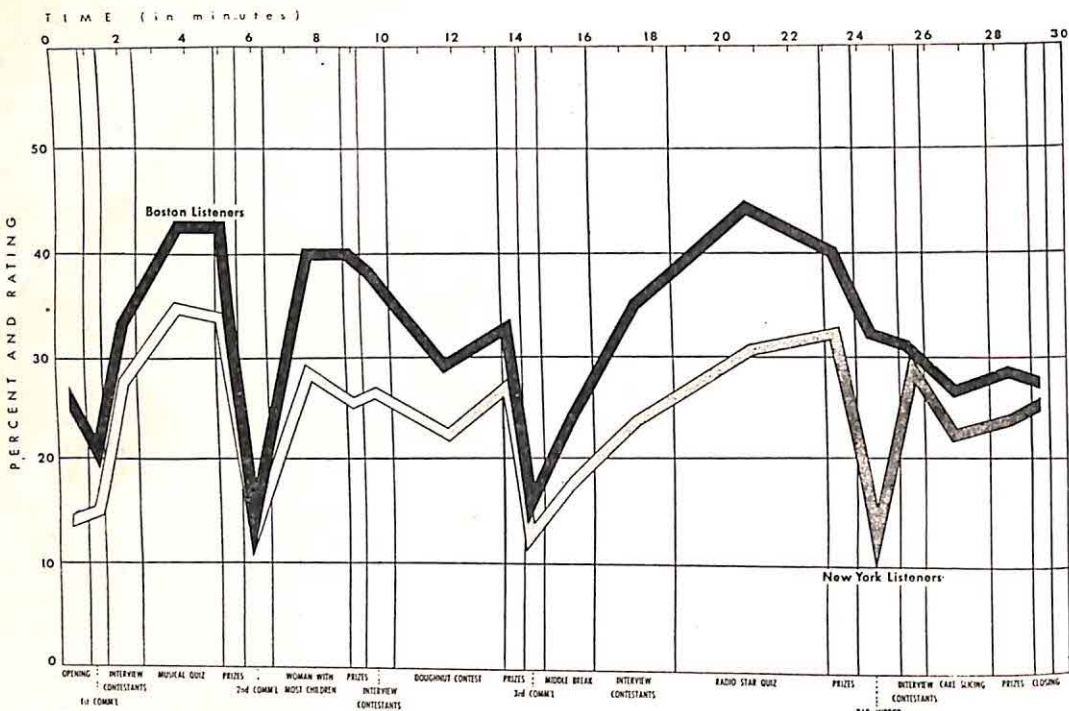


FIG. 4. Comparison of reaction trends for Boston and New York listeners, Boston audience participation program.

having been one of the items adding most to their enjoyment. Self-testing rated the highest of any element for satisfied, conditional and dissatisfied listeners in both New York and Boston samples. More than two-thirds of each sample derived a great deal of enjoyment from "hearing people have a good time" (described as empathy in Table 16). On the other hand, the two aspects of the program that had the least appeal were the prizes and the contests.

The brunt of the success of any audience participation show rests largely on the M.C.

and it is interesting to note that the two samples reacted similarly to the M.C. of this program, although again the Boston audience reacted somewhat more favorably probably because of its greater familiarity with his personality prior to the time of the Program Analyzer test. In Table 15, four aspects of the M.C. are considered; the per cent of each sample enjoying each is indicated. The M.C.'s humor and jokes were least appreciated, but even so, nearly three-fifths of the Boston audience found them "very enjoyable."

Table 13

Per Cent Satisfied, Conditional, and Dissatisfied Listeners to the Boston Program

	New York Sample	Boston Sample
Satisfied	46%	51%
Conditional	20	32
Dissatisfied	34	17
	100%	100%

Table 14

Gratifications and Appeals of the Boston Program

	New York Sample	Boston Sample
Self-testing	77%	79%
Empathy	69	68
Human interest	42	59
Information	46	58
Humor	52	56
Prizes	31	44
Contests	31	41

Table 15

Favorable Opinions about Personality and Performance of the Master of Ceremonies

	New York Sample	Boston Sample
M.C.'s personality	67%	79%
M.C.'s handling of contestants	65	79
M.C.'s voice and manner of speaking	67	74
M.C.'s humor and jokes	46	57

Summary

The results of the New York-Boston tests point to the advisability of testing "local" programs with local audiences if accuracy for general level of response and program appeal is the principal question. On the other hand, the program's producer could ascertain the most important strong and weak points of the program just as well from the New York sample as from the local Boston sample, inasmuch as the high- and low-spots tended to parallel each other up to the very closing of the broadcast. We are of the opinion that this would generally be the case except for programs that are purely local in their content, orientation and atmosphere. Other Program Analyzer tests have established, for example, that

quiz questions of the local sort, that cannot be answered other than by local audiences, have little or no appeal for outsiders. If self-testing were not the main appeal of quiz programs, then this might not make such a difference, but self-testing has repeatedly been found to be the principal appeal.

The results of the New York-Los Angeles tests demonstrate that national network programs can be satisfactorily analyzed, at least for urban audiences, with samples of subjects drawn from different geographic areas. We are of the opinion that the Program Analyzer technique has a general usefulness for the analysis of such programs regardless from what urban area the sample may be drawn. Unfamiliarity of listeners with the principal character (or characters) of a program will tend to affect somewhat the general level of listeners' likes and dislikes, but not decisively so, particularly since the major high- and low-spots of a program evidently will be the same. From the point of view of diagnosing the appeal of a program as a whole and its various parts, as well as making recommendations for the improvement of future broadcasts, this is the primary consideration.

Received April 1, 1950.

Early publication.

The Effect of Color in Direct Mail Advertising

J. William Dunlap

Harvard School of Public Health

There have been a number of discussions concerning the value of color in direct mail advertising. Birren presents the results of several studies which indicate that color pulls more returns than black and white in this type of advertising.¹ The results of the several studies presented by Birren were, in some cases, slightly contradictory. Furthermore, pertinent data needed for determining the statistical significance of the differences were not presented. Among the colors Birren found to be best were: yellow, goldenrod, blue, and cherry-red. It was felt that a study should be conducted in which statistical tests could be applied to determine the significance of the results.

The colors tested in the present study were: yellow, blue, and cherry with white as the base or control "color." It was intended to use the psychological primary colors of red, blue, and yellow. However, since the true primary colors could not be obtained from the paper manufacturer, colors were chosen that were as close to the psychological primaries as was possible. The colors blue and yellow closely match the primaries, but the cherry is somewhat different.

It is necessary that the reader be given some idea of the relative "brightness contrast" between the print used and the color of the cards. Paterson and Tinker stated that legibility and speed of reading depend upon the "brightness contrast" between the print used and the background for the print.² The background colors for the black print used in this study are all comparatively light. However, these colors approach maximum chroma, or saturation. There is little difference in the "brightness contrast" between the white and yellow cards and between the blue and cherry cards. The contrast of the print on the blue

and cherry cards, however, is not as great as the contrast between the print and the white and yellow cards.

The nature of the "advertising" material was a card notifying members of the Kansas State Alumni Association that their annual membership was expiring and this was their opportunity to renew it.³ The cards were of three ply colored Hammermill Index Bristol. The dimensions of the cards were 3"×5". The message was printed in black on all cards, regardless of color.

During the first week of each month, cards were sent to all members whose membership would expire that month. Every fourth person on the mailing list received the same color or card. No person received more than one card.

Between the dates of Dec. 1, 1948 and August 6, 1949, a total of 572 cards were mailed out. The distribution was as follows: 147 white, 144 yellow, 141 blue and 140 cherry. Table 1 shows the number of cards of each color that were sent out each month and the number returned. The highest percentage of returns, 50.7, was for the yellow cards. The rank order of returns for the other colors was blue 46.1%, white 40.8% and cherry 38.6%.

The question now is whether the observed differences in returns are due to chance. Under the null hypothesis it is assumed that no difference exists due to the effect of color. A simple and direct overall test for this assumption is provided by chi square. Calculation gave a chi square of 2.9, which for 3 degrees of freedom is significant at about the $P = .40$ level. Thus the null hypothesis of homogeneity must be accepted, and it is concluded that no difference in "pull" was found due to the use of color.

This result does not support Birren's con-

³ This study could not have been conducted without the cooperation and support of Kenny L. Ford, Alumni Secretary, KSC. The writer is also indebted to Dr. Roy C. Langford for his suggestions and encouragement.

¹ Faber Birren. *Selling with color*. New York: McGraw-Hill, 1945.

² D. G. Paterson and M. A. Tinker. *How to make type readable*. New York: Harper and Brothers, 1940. Ch. 10, "Color of Print and Background," pp. 118-129.

Table 1

The Number of Cards by Color Sent Out and Returned by Mailings

Month Mailed	White		Yellow		Blue		Cherry	
	Sent	Ret'd	Sent	Ret'd	Sent	Ret'd	Sent	Ret'd
Dec.	8	1	7	4	7	2	6	2
Jan.	9	5	8	4	9	3	9	2
Feb.	17	12	16	12	17	14	17	12
Mar.	28	16	28	17	27	15	27	17
Apr.	18	3	18	8	16	9	16	5
May	25	6	25	11	24	4	24	2
June	10	2	10	1	9	1	9	2
July-Aug.	32	15	32	16	32	17	32	12
Total	147	60	144	73	141	65	140	54
% Ret'd	40.8%		50.7%		46.1%		38.6%	

tention that color affects "pull" in direct mail advertising. Birren reports a study done by a manufacturer, who tested colored envelopes for the number of responses and obtained the following results: blue 7.8%, yellow 6.8%, goldenrod 6.4%, green 6.0%, pink 5.8% and white 3.1%. He reported another study in which colors were measured in terms of percentage of orders they produced. The results given were: goldenrod 21.42%, pink 17.83%, green 17.82%, white 17.29%, kraft 15.89% and an old envelope (color not stated) 9.75%. Birren reports still another study done by a milling company who, testing "return cards," found that 50.6% of the returns were cherry-

red cards, while white and blue pulled 32.7% and 16.7%, respectively.

The available evidence as to the effect of color in direct mail advertising is contradictory. It is possible that the results from the alumni membership cards represent a sampling error due to the content of the cards, the personnel sampled, or to the size of the sample. Examination of all the data available gives the impression that black on yellow, buff, or goldenrod has a greater pull than does black on white. In view of the possible practical value of color in direct mail advertising the problem should be subjected to further investigation.

Received September 22, 1949.

Brand Discrimination among Cigarette Smokers

C. K. Ramond, L. H. Rachal, and M. R. Marks

Tulane University

It is a matter of common knowledge that the essence of cigarette advertising is the claim that the particular cigarette is distinguishable from other brands. Habitual smokers frequently comment that they are able to identify their own brand. If these claims be true, it follows that there must be discriminable differences among brands, and the problem of ascertaining the extent of such differences is one of interest to psychologists.

The writers have found only two studies which are immediately relevant. Husband and Godfrey (2) in 1934 worked with 5 different brands. They requested 51 *Ss* to attempt the identification of 4 cigarettes, under the condition that *S* was told only that his cigarette was included among the 4. The report does not state clearly whether the *S* knew just what brands were possible choices. *Ss* were blindfolded. The data are given in terms of percentage of correct and incorrect identifications for each brand tested. Although no statistical techniques were employed to evaluate the data, Husband and Godfrey concluded that most cigarettes tested were identified correctly slightly more times than would be expected by chance. They noticed anomalous findings, e.g., "Camels are identified as Chesterfields more often than as themselves" (2, p. 222).

There is evidence, however, that the use of a blindfold obscures the central problem. Hull (1) in 1924, while studying the physiological effects of tobacco smoking, found that his *blindfolded Ss* frequently could not distinguish between real tobacco smoke, and warm-moist air, when both were inhaled through a pipe mouthpiece. Those readers who are habitual smokers may recall that when they are smoking in the dark they are sometimes not sure of whether they are smoking at all!

The present investigation was designed to test capacity for discrimination among various popular brands of cigarettes, to wit, Camels, Chesterfields and Lucky Strikes, when *S* was allowed to see both the cigarette and the smoke.

Answers for the following questions were sought:

1. Do correct identifications exceed chance expectancy, and, if so, what is the margin of improvement over chance?

2. Do *Ss*, who are permitted to smoke the brands interchangeably, make higher identification scores than *Ss* who are required to smoke a single brand until they commit themselves as to its identity?

3. Does an *S* who habitually smokes a given brand identify that brand correctly more often than do *Ss* who habitually smoke some other brand?

Procedure

Subjects: Only smokers who consumed at least one pack of cigarettes in a four day period participated in the study. They were 200 in number, of both sexes, accidentally sampled from Tulane University students. *Ss* were approached outside of classrooms, in student centers, and on streets near the campus.

Apparatus: This consisted of mimeographed identification questionnaires, gummed labels, candy mints, and most important, 1,200 cigarettes—400 each of the brands named above. Plain gummed labels of identical size were pasted in identical positions over the brand names of the cigarettes used in the test-smoking. Each label bore an initial—X, Y, or Z. These non-committal identifying marks were coded to the brand names, and, for security purposes, the code was changed halfway through the study. The questionnaire was as follows:

Circle the brand name which you think the labeled cigarette is:

X	Y	Z
Camel	Camel	Camel
Chesterfield	Chesterfield	Chesterfield
Lucky Strike	Lucky Strike	Lucky Strike

Method: Two principal groups of *Ss* were employed. An attempt was made to have in

each group Ss who habitually smoked each of the brands tested, and also, Ss who habitually smoked other brands. Actually, in Group I, there were 24 Ss who smoked Camels, 27 who smoked Chesterfields, 24 who smoked Lucky Strikes, and 25 smokers of miscellaneous brands. In Group II, the corresponding breakdown was, 25 smokers for each of the categories described.

Practice Smoking: Instructions given to each S in Group I were as follows: "Light all three of these cigarettes (one each of the three brands) and smoke them *interchangeably*. Notice the brand which you are smoking and look for characteristics which will help you identify it later. Smoke until you feel you will be able to make the identifications." Instructions given each S in Group II were as follows: "Light only one of these cigarettes. Notice its name and smoke it until you feel that you can identify it later. Then take another cigarette and proceed in the same way. Put out one cigarette before lighting another."

Test Smoking: Each S was given a candy mint to clear the taste of the practice smoking. The method paralleled that used in the practice, but this time the names of the cigarettes were masked by the initialed gummed labels. The Group I Ss indicated their identifications on the questionnaires after smoking all three cigarettes interchangeably. The Group II Ss indicated a choice each time they completed one of the three cigarettes. In no case included in the data, did an S use a given brand name to identify more than one of the test cigarettes. None of the Ss was told of his degree of success.

Results

Table 1 shows the number of times each cigarette was identified as itself and as some other cigarette, by experimental groups.

The first pertinent question is, "Do correct identifications exceed chance expectancy, and if so, what is the margin of improvement over

chance?" The χ^2 test of independence of principle of classification was applied to the data of Table 1. With 4 degrees of freedom, $\chi^2=32.75$. A value as great as this could be expected by chance only 1 time in 100,000. The chance expectancy is of course 33.3%. The overall average percentage of correct identifications was 44.5. Thus, the margin of increase, while highly significant, is small in magnitude. The figures for individual cigarettes and individual groups do not differ significantly from the average figure of 44.5% correct identifications. The actual percentages were: For Group I, Camels, 48; Chesterfields, 40; Lucky Strikes, 41. For Group II, 40, 41, and 50, respectively. It appears that, while no individual cigarette is more "distinctive" than any other, there is a slight but significant discriminability among the three brands tested.

The second question is, "Do Ss who are permitted to smoke the brands interchangeably make higher identification scores than Ss who are required to smoke a single brand until they commit themselves as to its identity?" This question may be answered by comparison of Groups I and II. Group I averaged 44% correct; Group II averaged 45% correct. The difference is not statistically significant. It appears that this limited practice is not effective in differentiating between the groups. It is quite possible that the two kinds of practice did not change the scores at all, so that if a third group had been employed in which only the test smoking was administered the identification scores would have been as high as those obtained in this study. It should be noted that the Group II conditions approximate the "real life" situation. Habitual smokers smoke only one cigarette at a time!

The third question is, "Does an S who habitually smokes a given brand identify that brand correctly more often than do Ss who habitually smoke some other brand?" Of the Camel smokers, 75% were able to identify their own brand. Corresponding figures for Chesterfields and Lucky Strikes were 70% and 74%. There is no significant difference among these percentages. Smokers of particular brands tested (i.e., excluding the miscellaneous smokers) identified the popular brands other than their own less frequently

Table 1

Number of Identifications of Cigarette Brands by Brand Name and by Experimental Group

Actual Brand	Identified As					
	Camel		Chesterfield		Lucky Strike	
	I	II	I	II	I	II
Camel	48	40	24	33	26	27
Chesterfield	23	35	46	41	31	23
Lucky Strike	30	24	31	27	41	50

than did the habitual smokers of those brands. Of the combined Chesterfield and Lucky Strike smokers, only 39% identified Camels correctly; of the combined Camel and Lucky Strike smokers, only 42% identified Chesterfields correctly; of the combined Camel and Chesterfield smokers, only 44% identified Lucky Strikes correctly. There are no significant differences among these percentages, but all of them are significantly less than the percentages of correct identifications of "own" brands given above. Miscellaneous smokers did even less well. Camels were identified correctly by only 24%, Chesterfields by 14%, and Lucky Strikes by 22%. Again, there are no significant differences among these percentages, but all of them are significantly less than the percentages for habitual smokers of the brands tested. The answer to the question, "Do smokers know their own brands?", is fairly clear. Although the overall percentage of correct identifications was 43.5, the "own brand" identifications averaged 73%; identifications of popular brands by smokers of other popular brands averaged 42%; identifications by smokers of miscellaneous brands averaged 20%. There is thus a well-defined tendency for smokers of particular brands to know that brand. Smokers of miscellaneous brands score lower than could be expected by chance. They seem to have positive misinformation about the differential tastes of the popular brands.

Summary

One hundred Ss, each of whom customarily smoked at least one pack of cigarettes each four days, practice-smoked one each of three popular cigarettes—Camels, Chesterfields, and Lucky Strikes—until they felt that they could identify them. In practice and in test they were allowed to interchange the cigarettes, after lighting all three. Another 100 Ss, both in practice and in test, smoked only one of the brands until practice was over or identification made, then proceeded to the next, etc. In the test situations the cigarettes were masked by uniformly sized and positioned gummed labels. All Ss recorded on a questionnaire their identifications of the three brands.

Inasmuch as the sampling was predominantly of college students, generalization from the

results is applicable only to such a population except insofar as the sample tested reasonably represents the general smoking population. Further, in a short-run test such as this, such factors as throat-irritation or, contrariwise, acquired insensitization, are minimized. In a more extended smoking test the obtained frequencies of correct identifications might well change. With these restrictions in mind, the following conclusions were reached:

1. All three brands tested were identified correctly an average of 44% of the time as against a chance expectancy of 33.3%. The increase, though slight, is significant statistically.
2. There were no significant differences in frequencies of correct identifications among the three brands tested.
3. There was no significant difference in correct identifications between those Ss who smoked their cigarettes simultaneously and interchangeably, and those who were limited to one cigarette at a time. The latter group is that which corresponds more closely to actual smoking practice. It is suggested that the training used in this study was not sufficiently extensive or intensive to be evocative of differences in performance.
4. Habitual smokers were able to identify their own brand significantly more times than smokers of other brands; this facility was uniform in habitual smokers of all three brands tested.
5. It would appear that claims of cigarette advertisers and habitual smokers to the effect that there are discriminable differences among various brands are technically true, but actually of small magnitude.
6. No data of this study indicated greater discriminability of any particular brand, or greater discriminatory capacity by smokers of any particular brand.

Received October 20, 1949.

References

1. Hull, C. L. The influence of tobacco smoking on mental and motor efficiency. *Psychol. Monographs*, 1924, 33, 161.
2. Husband, R. W., and Godfrey, J. An experimental study of cigarette identification. *J. appl. Psychol.*, 1934, 18, 220-223.

Report on the Journal of Applied Psychology for 1949

A summary of the materials published in Vol. 33 of the *Journal of Applied Psychology* is presented in the following table together with the corresponding data for Volumes 27-32.

Year	Vol.	Articles	Book Reviews
1949	33	76	33
1948	32	78	34
1947	31	80	29
1946	30	67	30
1945	29	53	18
1944	28	54	18
1943	27	58	9

The number of pages printed from 1946 through 1949 is as follows:

	1946	1947	1948	1949
No. pp. ("reg.")	476	478	485	479
No. pp. ("early")	192	186	199	138
Total	668	664	684	617

Vol. 33 was one page shy of the budgeted limit of 480 pages. Early publication contributed a total of 18 extra articles and 138 extra pages for our APA reader-owners and outside subscribers.

Early publication in 1949=18 articles; in 1948=21 articles; in 1947=28 articles; in 1946=16 articles. The lag in publication of "early publication" articles ranges from 2 to 4 months with a median of 3 months.

In addition to printing "date of receipt" at the end of each article, "early publication" is also added if such be the case. Thus all information concerning editorial action on published articles is available to contributor-owners, to reader-owners, and to outside subscribers.

The disposition of manuscripts received during 1949 and in each of the three preceding "post-war" years is as follows:

	1946	1947	1948	1949
Accepted	67	59	89	102
Rejected	38	44	35	74
Total Received	105	103	124	176
Per Cent Rejected	36	43	28	42

The number of Mss. received shows a steady post-war increase throwing a constantly heavier burden on the editor and his Consulting Editors.

A larger number of manuscripts would have been rejected had it not been for our policy of giving authors a chance to revise manuscripts in accordance with detailed suggestions provided chiefly by our Consulting Editors. Of 102 Mss. accepted in 1949, 61 were accepted "as is" and 41 were accepted following revision by the author.

The lag in publication of articles published in regular turn varies from time to time because manuscripts are not received in an even flow. During 1949 the low point was 5 in December and the high points were 23 in April and 23 in October. The median lag for regularly scheduled articles in Feb., Apr., and June 1947, for the same months in 1948, for the same months in 1949, for Aug., Oct., Dec., 1949, and for the same months in 1950 is as follows:

Median Months of Lag

Feb. Apr. 1947	Feb. Apr. 1948	Feb. Apr. 1949	Aug. Dec. 1949	Feb. Apr. 1950
12	9	7	9	10

Lag in publication is thus shown to be increasing since the first half of 1949. As of April 15, 1950 there are 39 accepted Mss. on hand plus 14 Mss. accepted if revised, and 7 Mss. "action pending," or a total of 60 Mss. likely to be published in August 1950 or later. The estimated lag for the last few Mss. accepted during the first half of April 1950 is 12 months.

The problem of lag in publication is thus once again becoming acute in spite of the continued policy of: 1. "Brevity consistent with clarity." 2. Early publication at author's expense. 3. Use of American Documentation Institute for large, unwieldy, and costly tables and figures.

The APA Council of Representatives *Directive* adopted at the Denver meeting in September 1949 will aid the editor in persuading authors to abbreviate their reports and to make increased use of ADI. During 1949, seven articles used ADI. During the first half

of 1950, six articles used ADI. Vigorous action by the editor is the only way lag in publication can again be reduced to a desirable six or seven month lag. If the postwar pressure continues to increase and we have exhausted all possibilities of "brevity consistent with clarity" and use of ADI then the Board of Editors and the Committee on Publications will be forced to consider an increase in the budgeted number of pages per volume.

The problem of the long delay between submission of edited copy to the printer and the receipt of a given issue by subscribers continues to be acute. An increased number of authors write in concerning their reprint orders because of the long time elapsing between appearance of an article in a given issue and receipt of reprints.

The new cover page and the new double column format has been received with favor judging from unsolicited comments. The 480 pages per volume of the old format with about 500 words per page has been reduced to 384 pages per volume in the new format but with about 750 words per page. Thus, there will be a slight increase in the number of articles that can be published per volume beginning with 1950.

The editor again wishes to express his appreciation to his Consulting Editors for invaluable aid in evaluating manuscripts and in submitting detailed suggestions for revision when such is indicated.

Donald G. Paterson
Editor

Received April 22, 1950.

Book Reviews

Drake, Frances S., and Drake, Charles A. *A human relations casebook for executives and supervisors*. New York: McGraw-Hill Book Company, Inc., 1947. Pp. xiv+187. \$2.50.

In reviewing a recent textbook on personnel management I was impressed by the thoroughness of its presentation of facts, principles, and techniques but felt that it was very dry reading. The structure was there, all neatly organized and patterned, but it lacked the living tissue necessary to make it come alive.

This "human relations casebook" tries to add flesh and blood to the formal structure of personnel management and does it pretty well. It presents 75 case histories, drawn from actual experience and descriptive of a wide variety of situations in industry. Section I, Adjusting the Human Resources, includes 17 cases involving selection, transfer, emotional deviations, training, retirement, etc. Section II, Developing Attitudes and Sentiments, presents 11 cases illustrative of factors affecting employee morale. Section III, Using and Abusing Incentives, portrays 11 incidents dealing with wage incentives and non-financial rewards. Section IV, Bargaining with Individuals and Groups, presents 8 cases involving handling of grievances and labor demands. Section V, Mobilizing the Brain Power of an Organization, deals with the use of conferences, suggestion systems, and work simplification programs through discussion of 6 cases. The final Section VI, the Ways of Executives and Supervisors, portrays 22 specific incidents illustrative of good and poor personal traits and behavior patterns of managers.

Although the case histories are quite brief and intentionally "stripped of much background and emotionally toned descriptive matter," they focus attention on the principles involved. To aid the reader in profiting vicariously from the successes and errors of others, each case is presented in the following manner: first, the area illustrated by the case is briefly outlined; second, the case history is presented; third, interpretive comments are made, generalizing from the specific incident; fourth, questions for group discussion are

raised; fifth, the reader is asked to formulate in his own words the primary lesson or principle drawn from the case under discussion. Facilitating the use of the book as a text for training group or classroom discussion are well selected and annotated bibliographies at the end of each of the six major sections.

Learning (or teaching) through case examples has obvious limitations. Solutions successful in one situation frequently fail in another and it is a common human error to generalize too quickly from a few instances. The authors seem to be aware of this in their deliberate cutting out of many of the details of the cases and in their emphasis upon the interpretation of the situation in terms of general principles. The principles do not grow out of the cases so much as the cases are illustrative of the principles. Looked at in this light, this reviewer felt it inappropriate to make what seemed on first reading to be obvious criticisms on the briefness of the cases and the failure to include certain expected types of situations.

The real merit of this book lies in the soundness of the particular principles illustrated, the value of the case examples in stimulating thought by the reader, and its general usefulness as a supplementary textbook or as a workbook in a supervisory development program.

Albert S. Thompson

*Teachers College,
Columbia University*

Mossin, A. C. *Selling performance and contentment in relation to school background*. New York: Bureau of Publications, Teachers College, Columbia University, 1949. Pp. viii+166. \$2.75.

Do salesgirls who have completed high school courses in "distributive" subjects perform more efficiently, and are they more contented with their work than salesgirls who have not had such courses? These are the two major questions that this study posed for investigation.

The subjects were 94 salesgirls from a large department store in New York City, rather

homogeneous with regard to such factors as period of employment, marital status, and duration of high school attendance prior to employment. They differed in having followed in high school what the author classified as either a college preparatory, commercial, distributive occupational, or clothing arts curriculum.

Job performance was measured by means of subjective criteria derived from independent ratings by four trained shoppers who observed each girl at work. Job contentment was evaluated by responses to three instruments specifically constructed for this investigation, namely a Job Functions Interest Blank, a Job Conditions Satisfaction Questionnaire and a Job Ranking Test.

No significant differences were found between the performances of salesgirls with different high school curricular backgrounds. Some slight tendencies were evidenced for girls who had taken the "distributive" curriculum to score higher on measures of job contentment than did girls from the other curricular groups. The "distributive" group expressed a greater desire to remain in sales work than did the three other curricular groups.

The study thus lends some support to other investigations which have indicated that interests, although not always useful as predictors of job performance, are indicative of motivations important to job contentment, worker morale, and reduction of employee turnover. However, certain weaknesses in experimental design forced upon the investigator by the particular department store situation tend to vitiate positive conclusions from the data.

The salesgirls were drawn from 41 different departments of the store, thus making comparisons between their selling tasks rather tenuous in view of differences in materials sold and problems encountered in selling. Only 13 of the 94 subjects were classified in the "distributive" curricular group and six of these had taken only two "distributive" courses. The author had to rely upon subjective criteria for the evaluation of job performance. The reliabilities of these (ratings by shoppers) were quite low and correlated .30 or lower with supervisors' ratings of the same employees.

Perhaps an inherent limitation in this ap-

proach to comparing workers' high school backgrounds is that these do not in themselves reveal attitudes toward curricula and work that are indicative of the motivations underlying vocational choice. We must know more about the meaning of vocational choice to individual students and workers if we are to evaluate properly the contributions of their backgrounds to job performance and contentment.

Daniel Raylesberg

B'nai B'rith Youth Organization

Vernon, Phillip E., and Parry, John B. *Personnel selection in the British Forces*. London: University of London Press Ltd. 1949, 324 pp. 20/net.

This book presents a summarized account of the application of psychological methods to personnel selection in the British Navy, Army, Air Force, and Army Territorial Service during World War II. The book is addressed to both industrialists and educators with the hope that the methods employed during the war in the British Armed Forces may be found useful in the peace-time selection of employees and in student selection and guidance. Throughout the book the authors have attempted to present their material as an integral part of the whole field of personnel psychology. War-time experience in the application of psychological methods to selection problems is, therefore, integrated with the knowledge which existed in the field prior to the war. Only material pertaining to personnel selection is included; applications of psychology to training, design of equipment, morale and so forth are not treated.

Part I is concerned with the organization of selection programs, the general procedures employed and the work of psychologists in the Royal Navy, Army, Royal Air Force, and Army Territorial Service. The work of non-military psychologists is not described. A chapter on the rise of vocational psychology is also included. This part is designed to give the reader a background for understanding the organizations within which psychologists worked. Unless this part is read first it will not be easy to read the later chapters, particularly since many abbreviations are employed.

Appendix I, Abbreviations, can be used for reference purposes.

In part II the authors present the highlights regarding the applications of psychology to personnel selection in the British Armed Forces. The first two chapters of this section are excellent condensed statements of the basic principles of vocational classification and the value of personnel selection procedures as a whole in the Armed forces. Other chapters are concerned with the biographical questionnaire, the interview, principles of psychological testing, and the various types of tests—intelligence and educational, non-verbal, mechanical, special aptitude, and temperament tests. A separate chapter is devoted to selection findings in the Royal Air Force. A chapter on conclusions completes Part II.

In the judgment of this reviewer the authors have made three significant contributions in the writing of this book. First they have presented a brief, clear description of selection procedures as they existed in the British Armed Forces during World War II. The student of military and personnel psychology will welcome having this material in a readily available form. Second, the book provides a good summary of the applications of psychology to personnel selection before the war. The chapters on the biographical questionnaire, the interview and principles of psychological testing are particularly good in this respect. Third, the results obtained during the war are critically evaluated and integrated with the knowledge which existed prior to the war. In a sense, the book almost impresses one as a textbook in personnel psychology rather than being a description of war experiences with the applications of psychology to personnel selection. The book is by no means limited to a description of findings. The contributions of war experiences to the total field of personnel psychology and the peace-time applications of these experiences are continually kept in mind by the authors.

As one reads this book he is impressed with the similarity of findings in the British and American Armed Forces. The types of tests which "worked" were much the same, the interview was beset with identical difficulties and the criterion was a problem for the British as well as American psychologists. Thorough

follow-up studies were equally few in number in the armed forces of the two countries.

The serious student of personnel psychology may be disappointed to find that data are presented only in summarized form. It is stated by the authors that they hope more detailed reports can be published elsewhere. The book is well written for its intended audience and will be widely read by personnel psychologists, personnel administrators and guidance workers. It will be useful as a reference work or supplementary text in courses concerned with personnel selection.

Dewey B. Stuit

State University of Iowa

Warner, L. W., Gardner, B. B., Henry, W. E., and Haggard, E. A. *Identifying and developing potential leaders*. New York: American Management Association Personnel Series No. 127. 1949. Pp. 26. \$0.75.

This issue contains a series of four papers presented at the AMA Mid-winter Personnel Conference with the sub-title "Social Science and the Management of American Business—A Report to Management from the University of Chicago."

The participants, and the titles, were: W. Lloyd Warner: "Individual Opportunity—A Challenge to the Free Enterprise System"; Burleigh Gardner: "Conserving and Developing our Human Resources"; W. E. Henry: "Identifying the Potentially Successful Executive"; and E. A. Haggard: "Social and Psychological Factors in Work Adjustment."

While it well may be that American industry is satiated or even confused by an inundation of "challenges," these associates of the University of Chicago's Committee on Human Development unhesitatingly throw out several more. Warner describes the uncomfortable situation that the usual social mobility routes upward (occupation and education) to higher status positions are becoming inadequate and indeed are closed in many areas. The resultant is a sort of a mass frustration, a willingness to give up the old system of individual mobility, and a readiness to blame the system.

Gardner discusses the overall problem of loss of satisfaction and deteriorating morale as brought about through such organizational

problems as technological change, "bigness" and over-extended hierarchy, decline of small, successful, satisfying, owner-managed firms, and the effects of specialization. Paradoxical to the apparent purpose of the papers (as revealed in the publication title) is "See that the good men in your organization are found and give them a chance to use what they have" (p. 12) (!).

Henry brings out the problem of the individuals who run the organization—the leaders. He ascribes the qualities of executive leadership as being to a great extent the qualities of individual personality. Personality research, according to Henry, among successful men in business and industry has revealed a pattern of common personality characteristics and that this is a pattern of fairly long standing. Executive skills may be equal but the fundamental personality organizations may be such that one executive is successful and another isn't. He regards the executive as a "particular kind of personality" worthy of personality analysis as used elsewhere. For illustration he indicates that the successful executive thinks in terms of the job hierarchy continuously, and has a positive reverence for a competent parent image. But in general he casts research questions rather than gives leads to identify potentially successful executives.

Haggard relates the above points to the necessity of satisfying the run-of-the-mill workers. He is particularly critical of "mechanized selection" by means of test scores where the factors of interests, motivations, and personality structure are not estimated at the hiring. (He speaks as if most workers got employed because of test scores whereas only about 15% of U. S. companies use any test at all.) He appeals for greater management recognition of human relationships on the job, of emotional needs, of the complete individual personalities. Again, "We need to locate the men who can rise in the organization, and help them do so" (p. 21).

All this is important and interesting. But it doesn't add up to "identifying and developing potential leaders." The audience undoubtedly was stimulated to support research,

activities, and efforts toward this end—which is all to the good.

Ralph R. Canter, Jr.

University of California at Berkeley

Robert Hoppock. *Group guidance; Principles, techniques, and evaluation*. New York: McGraw-Hill Book Co., 1949. Pp. 393. \$3.75.

Group Guidance is not a book for psychologists. As Dr. Hoppock states in his preface, "this book has been written for the beginner (teacher) who has been assigned some responsibility for group guidance and who wants to know what it is all about." To those who object to the term *group guidance*, as does the reviewer, Hoppock states, "The author of this book has no particular desire to debate the issue. . . . The author prefers the term 'group guidance' because it is short, concise, and descriptive and because it is extensively used in guidance literature. . . ."

The book should be a useful one to teachers and administrators in secondary education. It is pitched well as to level, it is readable, it contains many answers to "how" questions, and it makes none of the naive claims which have marred so many similar publications. Hoppock begins with brief definitions of his terms but, wisely, does not spend pages rehashing the history of the guidance movement. He relates the instructional, group approach to the individual counseling process objectively and well. He states the functions of the group approach in the first chapter and adheres to them for the rest of the way.

Part II, *Techniques*, is straightforward and clear. He considers the techniques of follow-up, visits to jobs and institutions, group conferences, student survey of jobs, case conferences, laboratory investigations, self measurement, and a number of less important methods. The reviewer takes general issue with the author only in the realm of *self measurement*. In justice to Dr. Hoppock it must be stated that his approach is a cautious one, although sad personal experiences leave some reviewer-doubts as to how cautious all readers of *Group Guidance* will be with test scores.

Part III, *Evaluation*, is excellent. Dr. Hoppock brings in pertinent research and applies it to his thesis in language which should hit the bulls eye for his intended audience. This section should be required reading for those students of professional education who believe that anything is possible by talking *at* groups of people. It will also be helpful to whose students of clinical psychology who believe that all worthwhile results must come from dealing with the individual.

Part IV, *Appendixes*, is a sound contribution to teaching methodology. Dr. Hoppock has been very successful in using the methods which he describes in these appendixes and reading them convinces the reviewer that most of us can learn from what he has presented.

A sound, well written book, objective in presentation, and cautious in approach which should be received well in educational circles.

Milton E. Hahn

University of California, Los Angeles

New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor,
Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota

- An index of nomograms.* Douglas Payne Adams, Editor. New York: John Wiley and Sons, Inc., 1950. Pp. 174. \$4.00.
- With brushes of comet's hair.* Cornelia H. Bogert. New York: Exposition Press, 1950. Pp. 165. \$5.00.
- A history of experimental psychology.* Second edition. Edwin G. Boring. New York: Appleton-Century-Crofts, Inc., 1950. Pp. 777. \$6.00.
- Readings on modern methods of counseling.* Arthur H. Brayfield. New York: Appleton-Century-Crofts, Inc., 1950. Pp. 566. \$5.00.
- College psychology.* Warner Brown and Howard C. Gilhousen. New York: Prentice-Hall, Inc., 1950. Pp. 485. Cloth, \$3.75; Paper, \$2.85.
- Marriage analysis.* Harold T. Christensen. New York: The Ronald Press Co., 1950. Pp. 510. \$4.50.
- Experimental designs.* William G. Cochran and Gertrude M. Cox. New York: John Wiley and Sons, Inc., 1950. Pp. 454. \$5.75.
- Abnormal psychology and modern life.* James C. Coleman. Chicago: Scott, Foresman and Co., 1950. Pp. 600. \$4.50.
- Readability.* Edgar Dale, Editor. Chicago: National Council of Teachers of English, 1949. Pp. 44. \$.60 per copy. \$.50 each for 10 or more.
- Child guidance approach to juvenile delinquency.* Eugene Davidoff and Elinor S. Noetzel. New York: Child Care Publications, 1950. \$4.50.
- Proceedings of second annual meeting of Industrial Relations Research Association.* Milton Derber, Editor. New York: Industrial Relations Research Association, 1949. Pp. 299. Annual subscription to IRRA Publications \$5.00.
- Rating employee and supervisory performance.* M. Joseph Dooher and Vivienne Marquis, Editors. New York: American Management Association, 1950. Pp. 192. \$3.75.
- The organization of mental abilities.* Jerome Edward Doppelt. New York: Bureau of Publications, Teachers College, Columbia University, 1950. Pp. 86. \$2.10.
- Guidance services in smaller schools.* Clifford P. Froehlich. New York: McGraw-Hill Book Co., Inc., 1950. Pp. 352. \$3.75.
- Guidance testing.* Clifford P. Froehlich and Arthur L. Benson. Chicago: Science Research Associates, 1949. Pp. 112. \$1.00.
- Psychology.* Henry E. Garrett. New York: American Book Co., 1940. Pp. 323. \$3.00.
- Fundamental statistics in psychology and education.* Second edition. J. P. Guilford. New York: McGraw-Hill Book Co., Inc., 1950. Pp. 633. \$5.00.
- Educational psychology.* Edwin R. Guthrie and Francis F. Powers. New York: The Ronald Press Co., 1950. Pp. 524. \$4.00.
- Counseling adolescents.* S. A. Hamrin and Blanche B. Paulson. Chicago: Science Research Associates, 1949. Pp. 380. \$3.50.
- The handbook of child guidance.* Ernest Harms, Editor. New York: Child Care Publications, 1950. Pp. 751. \$6.00.
- Industrial psychology.* Thomas Willard Harrell. New York: Rinehart and Co., Inc., 1949. Pp. 462. \$4.25.
- Personality, development and assessment.* Charles M. Harsh and H. G. Schrickel. New York: The Ronald Press Co., 1950. Pp. 518. \$5.00.
- The organization of behavior.* D. O. Hebb. New York: John Wiley and Sons, Inc., 1949. Pp. 335. \$4.00.
- Situational factors in leadership.* John K. Hemphill. Columbus: Bureau of Educational Research, Ohio State University, 1949. Pp. 136. \$3.00, cloth; \$2.50, paper.
- A miniature textbook on feeble-mindedness.* Leo Kanner. New York: Child Care Publications, 1950. Pp. 33. \$1.25.
- Gestalt psychology, its nature and significance.* David Katz. New York: The Ronald Press Co., 1950. \$3.00.
- Mental tests in clinics for children.* Grace H. Kent. New York: D. Van Nostrand Co., Inc., 1950. Pp. 180. \$2.45.
- The biology of human starvation.* Ancel Keys, et al. Minneapolis: University of Minnesota Press, 1950. Pp. 1360. \$24.00.
- The yearbook of psychoanalysis.* Sandor Lorand, Managing Editor. New York: International Universities Press, Inc., 1950. Pp. 317. \$7.50.
- The first two decades of life.* Frieda K. Merry and Ralph V. Merry. New York: Harper and Brothers, 1950. Pp. 581. \$3.75.
- Juvenile delinquency, modern society.* Martin H. Newmeyer. New York: D. Van Nostrand Co., Inc., 1950. Pp. 335. \$3.75.
- Social psychology.* Theodore M. Newcomb. New York: Dryden Press, 1950. Pp. 800. \$4.50.
- Child development.* Willard C. Olson. Boston: D. C. Heath and Co., 1949. Pp. 430. \$4.00.
- An introduction to therapeutic counseling.* E. H. Porter. Boston: Houghton Mifflin Co., 1950. Pp. 223. \$2.75.
- Human ability.* C. Spearman and L. L. Wynn Jones. New York: The Macmillan Co., 1959. Pp. 198. \$2.50.
- Criminology.* Revised edition. Donald R. Taft. New York: The Macmillan Co., 1950. Pp. 704. \$5.50.
- Management behavior and foreman attitude.* David N. Ulrich, Donald R. Booz, and Paul R. Lawrence. Boston: Graduate School of Business Administration, Harvard University, 1950. Pp. 56. \$7.50.
- Report of the Proceedings of the Second International*

- Congress for the Education of Maladjusted Children.* I. C. van Houte and Berthold Stokvis, Editors. Amsterdam, Holland: Systemen Keesing, Ruysdaelstraat 71, 1950. Pp. 448. \$4.50.
- Social class in America.* W. Lloyd Warner, Marchia Meeker, and Kenneth W. Eells. Chicago: Science Research Associates, 1949. Pp. 274. \$4.25.
- Blindness.* Paul A. Zahl. Princeton: Princeton University Press, 1950. Pp. 576. \$7.50.
- Human behavior and the principle of least effort.* George Kingsley Zipf. Cambridge: Addison-Wesley Press, Inc., 1949. Pp. 573.
- Directory of guidance agencies.* Ethical Practices Committee of the National Vocational Guidance Association. St. Louis: Dr. Nathan Kohn, Washington University, 1950. \$1.00.
- Revere Safety Test.* Revere Copper and Brass, Inc. Chicago: Science Research Associates, 1949. Test, 20 pages. Handbook, 60 pages. Test, \$.30; Handbook, \$.60.

Subscription Lists of the American Psychological Association

MEMBERS AND AFFILIATES

Approximately 9,800 names

The American Psychological Association maintains an address list of its members and affiliates, which is for sale providing the nature of its use is in conformity with the purposes of the Association.

1950 Prices

Envelopes addressed	\$35.00
(advertiser furnishes envelopes and pays express charges)	
Addresses on tape, not gummed	\$17.50
(suitable for a mailing machine)	

STATE LISTS

Priced according to number of names wanted

SUPPLEMENTARY LISTS

Approximately 3,500 names in total list
Individual journal lists vary from 400 to 1,600

The Association also maintains a list of subscribers who are not members of the Association (universities, libraries, industrial laboratories, hospitals, other types of institutions, and individual subscribers). The general list for all journals includes all types. Each single journal has a more specialized circulation.

For any one journal, envelopes addressed	\$15.00
For any one journal, addresses on tape	\$10.00
For all journals, envelopes addressed	\$20.00
For all journals, addresses on tape	\$15.00

For further information, write to
American Psychological Association
1515 Massachusetts Avenue Northwest
Washington 5, D. C.

Journal of Applied Psychology

VOL. 34, No. 5

OCTOBER, 1950

An Aptitude Test for Veterinary Medicine *

William A. Owens

Iowa State College

Particularly since the war, schools of veterinary medicine have been able to enroll for training only a relatively small percentage of their applicants. An increased emphasis upon the importance of identifying the best-qualified candidates has resulted in a careful reexamination of selection procedures and in recognition of the possible utilization of some sort of veterinary aptitude test.

Accordingly, the problem of the present investigation was to discover or to develop an efficient predictor, or predictors, of scholastic success during the first professional year of veterinary training.

Preliminary findings are based upon the records of all ($N=133$) freshmen and sophomores who were enrolled in the School of Veterinary Medicine at The Iowa State College during the academic year 1947-48.

Validation findings are based upon the academic records of 150 pre-veterinarians tested and *subsequently* enrolled in Veterinary Medicine at either Cornell University ($N=25$), Michigan State College ($N=41$), Kansas State College ($N=49$), or Iowa State College ($N=35$) during the academic year 1948-49.

Method

Indices already available were examined as to their predictive utility. These included: high school academic average, pre-veterinary college average, grade in certain specific pre-veterinary courses, scores on the ACE psychological examination, and sub-test scores on Form 20 of the Moss Aptitude Test for Medical Professions (3).

Since none of the above proved to be a highly satisfactory predictor of academic

* The writer wishes to acknowledge the invaluable counsel and assistance of Dr. James E. Wert.

success, it was decided to construct four special purpose tests. Two of these were 50-item achievement tests, constructed with the assistance of the departments concerned, and over the content of the two most predictive pre-veterinary courses—chemistry and zoology. The remaining two were, respectively, 60 and 50-item aptitude tests designed to measure the same abilities as the most predictive pair of sub-tests in the Moss ATMP. The first, called "Paragraph Comprehension," is a reading test; and the second, designated as "Verbal Memory," involves the timed study of standard selections with a subsequent objective test upon accuracy of recall. Both are entirely new, and their content was judged by members of *that* staff at Iowa State to be representative veterinary content. The usual psychometric procedures relative to establishment of time limits, analysis of items, and general revision, were employed following a cross-sectional experimental study of the four tests in 1947-48 and prior to a longitudinal study of their validity in 1948-49. The chemistry and zoology achievement tests were combined subsequent to the 1947-48 study, since their inter-correlation approached their reliabilities; the composite, shortened to 80 items, was simply called "Pre-veterinary Achievement." It may be noted that none of *these* test results have been employed in selection and that the data are, in this regard, uncontaminated.

In the validation study, the criterion adopted for the evaluation of veterinary aptitude was that of academic success during the first semester or first two quarters of professional training. A total grade-point average for this period was employed for three reasons: (1) such an average is highly correlated with ultimate scholastic standing; (2) academic

mortality is heaviest at this time; and (3) it seemed to be the longest waiting interval in a longitudinal study compatible with the urgency of the case. As these data were received from the several cooperating schools, each subject's grade-point average was assigned a standard score value in a distribution for the institution from which it had been obtained.

Statistical analysis of these data follows the usual correlational form with the exception that a discriminant function style of analysis has been employed in obtaining estimates of the weights to be assigned to each of the several new tests in order to produce composite scores which maximize the differences between certain performance groups. In addition, multiple biserial correlations have been employed at one point, after a method described by Wert (4), to estimate the combined effects of the tests in predicting the dichotomous criterion of performance.

Results

The primary results of this investigation have been summarized in four tables. Prior to some systematic comment upon them, it seems appropriate to observe that the "corrected" odd-even reliability of the composite or total test score derived from the several weighted sub-test scores was 0.88. This estimate is based upon results obtained from the entire tested population of 424 candidates for admission into veterinary training at the previously named cooperating institutions. Total test reliability thus seems reasonably satisfactory.¹

To proceed, in Table 1 are shown the correlations between various predictors and freshman veterinary average. Under "Existing Indices" it is interesting to note that pre-veterinary chemistry average appears to be the best single predictor. A possible artifact involved is the relatively large variance in chemistry grades as compared with that typical of other pre-veterinary subjects.

Sub-tests three and six of the Moss ATMP suggested functions to be measured by the two new aptitude tests, and the former may be broadly considered as models for the latter.

¹ As given with unlimited time in the preliminary study, the reliabilities for the sub-tests are Pre-veterinary Achievement, 0.65; Paragraph Comprehension, 0.78; and Verbal Memory, 0.74.

Table 1

Correlations of Various Predictors with Freshman
Veterinary Average
(Preliminary Study, N = 133)

Variable	<i>r</i> *
Existing Indices	
Total pre-veterinary average	.40
Pre-veterinary chemistry average	.47
Pre-veterinary zoology average	.41
Raw score on ACE	.02
Moss ATMP Subtests	
Visual Memory	.13
Memory for Content	.13
Comprehension and Retention	.22
General Information	-.11
Vocabulary	-.01
Understanding of Printed Material	.38
Application of Principles	.17
Logical Reasoning	.02
Four New Tests	
Chemistry Achievement Test	.27
Zoology Achievement Test	.12
Paragraph Comprehension	.47
Verbal Memory	.57
Sum of Paragraph Comprehension and Verbal Memory	.62

* At 5 per cent level $r = .17$; at 1 per cent $r = .22$.

As previously indicated, the "Four New Tests" were reduced to three following the evidence derived from the preliminary study and thru the combination of the chemistry and zoology achievement tests to form the revised Pre-veterinary Achievement Test. In spite of its apparently poor validity, this content was tentatively retained because it was recognized that the selection process at Iowa State, involving heavy emphasis upon pre-veterinary success in chemistry and zoology, brought about a tremendous restriction in range of talent on these tests which might not be duplicated at other institutions. In the event that it were not, it was felt that they could conceivably be of substantial value.

From Table 2, then, it is apparent that each of the three new tests much better predicts academic inability from a relatively low cutting score than academic standing from the total range of scores (this admittedly within the successful group). As evidence, the product-moment correlations in the left-hand column

Table 2

Correlations of Test Scores with Grade-Point Average (Validational Study, $N = 150$)

	P.-M Correlation*	Tetrachoric r (G.P.A.— $\frac{1}{2}$ vs. $\frac{1}{2}$)
Pre-veterinary Achievement (X_1)	0.24	0.41 (31%)
Paragraph Comprehension (X_2)	0.36 (0.47)	0.56 (25%)
Verbal Memory (X_3)	0.17	0.26 (28%)

* 5% $r = 0.16$ and 1% $r = 0.21$.

Table 3

Multiple Biserial Correlations and Discriminant Test Weights (Validational Study, $N = 150$)

	Multiple Biserial r 's (G.P.A.— $\frac{3}{4}$ vs. $\frac{1}{4}$)	Weights in the Discriminant Function
X_2 and X_3 (all students)	0.37	$v = 2.05 X_2 + X_3$
X_2 and X_3 (within schools)	0.43	
X_1, X_2 and X_3 (all students)	0.40	$v = -0.58 X_1 + 2.13 X_2 + X_3$
X_1, X_2 and X_3 (within schools)	0.45 (0.57)	

are consistently and substantially smaller than the tetrachoric correlations² in the right-hand column. These latter were obtained by arbitrarily dichotomizing grade-point averages at their mean, and by then dichotomizing test score distributions at the points of minimum error in classification, at or slightly above the lowest quartile point. The percentage to the right of each coefficient is that below the cutting score in the truncated tail of the test distribution in question.

Since it appeared to be most efficient in each case to set cutting scores on the tests near the 25th centile, it was arbitrarily decided to break the distribution of grade-point averages at the same level and to attempt to determine how best to weight each test to maximize the differences between these two segments of the criterion.

Thus, in Table 3, are shown the multiple biserial correlations and discriminant-function weights based upon this proposed dichotomy in the criterion. It may be noted that the correlations for "all students" are considerably smaller than are those "within schools." This is no doubt attributable to test-wise institutional differences so large that an individual's

scores might rank him in the highest quarter at school A and in the lowest at school B. The second series of discriminant function weights in Column 2 are those which were employed to derive a single series of composite or total scores, the "V scores," for all tests.

Table 4 is a summary table. In it appear the tetrachoric correlations between the dichotomized composite test scores and the dichotomized criterion. The per cent to the right of each coefficient, again, indicates the proportion of cases below the minimum error cutting score in the tail of the test distribution. Final results were cast in this form to make them coincident with the form of the practical question as it always arises, is a given subject above or below some critical test score, and how does this argue for his being above or below some accepted grade standard?

It is to be regretted that the number of cases for each institution is not larger, although application of the chi-square test to these data suggests that the *least* significant tabled relationship surpasses the 2% probability level. A sort of increase in numbers may, of course, be achieved by ignoring institutional differences and combining the data. This, naturally, lowers the apparent degree of relationship; in this instance it results in a single r_t of 0.52.

² Magnitudes estimated from the computing diagrams of Chesire et al. (1).

Table 4

Tetrachoric Correlations Between Composite Test Scores and Grade-Point Average (Validational Study)

	Cornell Univ. N = 25	Iowa State Coll. N = 35	Kans. State Coll. N = 49	Mich. State Coll. N = 41
High $\frac{3}{4}$ vs. Low $\frac{1}{4}$ (Criterion)	0.72 (16%)	0.70 (12%)	0.62 (18%)	0.48 (12%)

Interpretation

In interpreting these results several conflicting influences must be recognized. The estimates of test-criterion relationship provided in Table 4 may be thought of as overestimates for at least two reasons. First, they are based, after the fact, upon the most efficient test cutting score; whereas, in practice it may be impractical to find or to employ such a value, which will in any case show sampling fluctuations. Second, the scoring weights established for the combination of sub-test scores were derived from a composite of four populations and then applied to this *same* composite population. A "shrinkage" in discriminative efficiency must be expected when these weights are applied to the scores obtained from a new population, although the customary effect may be minimized in this instance since the original sampling was of a broader and more heterogeneous group than could have been obtained at a single school.

Running counter to these two influences, which would tend to make the values of Table 4 appear to be overestimates, is the undoubted fact that the coefficients shown have been depressed by a marked restriction in the range of talent existing *within* the validational group. At some institutions, the standard deviation of composite test scores is 25 to 30 per cent larger in the distribution of *candidates* for admission than in the distribution of *selectees*. This is a fact mainly attributable to current selection on the basis of pre-veterinary grade-point average—a test-correlated variable. If Kelly's (2) correction for homogeneity were applied to the present data to obtain estimates of test-criterion correlations within the population of candidates, many of the relationships here reported would be substantially increased in

magnitude. For example, in Table 2, the Paragraph Comprehension test would correlate 0.47 with the criterion instead of 0.36; and, in Table 3, the final multiple correlation would be 0.57 instead of 0.45.

In addition, there is evidence to suggest that within the groups admitted to veterinary training, those who *had* taken the tests, and who composed slightly more than half of the total number, were superior in performance to those who had *not* taken the tests. At least a partial explanation is that, almost without exception, those tested had had their pre-veterinary training at one of the four cooperating institutions. By and large students with this background make better grades and fail less frequently.

In evaluating these conflicting influences the writer is disposed to judge it as not unlikely that they approximately cancel and offset each other, and that the estimates of relationship reported are, therefore, not grossly in error.

Summary

A veterinary aptitude test has been devised which had the following characteristics in the populations studied:

- (1) It had a reliability of 0.88.
- (2) It had tetrachoric validity coefficients of from 0.48 to 0.72, against a grade-point average criterion.
- (3) It was a better predictor of the specified criterion than were pre-veterinary grades, singly or collectively, or the ACE.
- (4) It predicted most efficiently, within the validational group, from a relatively low cutting score.

Received October 24, 1949.

References

1. Chesire, L., Saffir, M., and Thurstone, L. L. *Computing diagrams for the tetrachoric correlation coefficient*. Chicago: The University of Chicago Bookstore, 1933.
2. Kelley, T. L. *Statistical method*. New York: Macmillan, 1923, pp. 223-228.
3. Moss, F. A., and Hunt, Thelma. *Aptitude test for medical professions*. Prepared especially for the War Department by the authors and under a committee of the Association of American Medical Colleges.
4. Wert, J. E. Unpublished manuscript available from the author, Iowa State College.

Minnesota Psycho-Analogies Test*

Abraham S. Levine

Human Resources Research Center, Lackland Air Force Base, San Antonio, Texas

This project was undertaken with the general objective of developing an evaluation instrument for psychology students. Since it was hypothesized that achievement in advanced psychology courses is to a large extent a function of a complex of general academic ability and previous psychological background, a test was developed to provide a composite measure of these factors. This test comprises items of the four-alternative multiple choice variety in analogy form. The first part of each item contains general vocabulary and information. The latter part of these items consists of a broad sampling of psychological terms, concepts and expressions which attempt to sample as widely as possible the content of all the major fields of psychology. The response alternatives are all psychological in content. Thus the first two terms of the analogy are of a general nature, usually non-psychological in content; while the third and fourth terms are psychological in character. An example of this type of item, with the correct response in italics, is as follows:

Orchestra: Violinist:: Test: (1. Battery, 2. Item Analysis, 3. *Item*, 4. Validity)

It was believed that a special analogies test of the type described would serve the dual function of selection and general evaluation, depending upon the point in a student's career at which it was administered. More specifically, it was anticipated that the test would serve some or all of the following purposes:

1. Selection of students for certain advanced courses or for certain sections of advanced courses in psychology.
2. Selection of graduate majors in psychology.

3. Selection of applicants for special training programs such as the Veterans Administration Clinical Psychology Program.
4. General evaluation of professional fitness of a student completing requirements for a degree.
5. Measure of growth of student by utilization of two equivalent forms of the test at different stages in his training.

Preliminary Editions of the Test

The first edition of the test, called "Psychological Analogies," was exploratory in nature. It consisted of 100 items. This test correlated .60 with combined midquarter and final examination scores for 92 students in a Senior College class in Vocational and Occupational Psychology at the University of Minnesota.

The second or preliminary edition of the test was given the title "Psycho-Analogies." It consisted of a total of 232 items, divided into two forms of 116 items each. While some of these items represented the more discriminating items of the original test, or modifications of these, the bulk of them were still new and untried. These new items were on the whole more carefully constructed, and greater attention was given to balancing subject matter content. The test was administered to several sections of Vocational and Occupational Psychology and Individual Differences at the University of Minnesota. These two courses are representative of intermediate and advanced courses in psychology in which juniors, seniors and graduate students are enrolled. A total of 161 cases was obtained on Form A and 156 cases on Form B. Correlation of a single form with combined midquarter and final objective examinations in the various sections ranged from .51 to .74. The available data also suggested that Psycho-Analogies was a slightly better predictor of course achievement than either a general analogies test (Miller Analogies, Form A or B) or a specially designed achieve-

* This article is based on the writer's Ph.D. thesis done under the direction of Prof. Donald G. Paterson and entitled "A Psycho-Analogies Test as an Evaluation Instrument for Psychology Students," completed in December, 1949, and on file in the University of Minnesota Library.

ment pretest. Moreover, a tendency was noted for the correlations between Psycho-Analogies and course examination scores to be higher in the more advanced sections.

In order to meet the need of a brief pretest for purposes of screening or sectioning certain psychology courses, a short form of Psycho-Analogies was developed, consisting of the 50 most discriminating of the 232 items. In those classes in which the test was not used as a basis for sectioning, so that the whole range of talent was available for purposes of validation, the correlations between the short form and combined midquarter and final examinations ranged from .41 to .68.

Minnesota Psycho-Analogies

The final edition of the test, Minnesota Psycho-Analogies, consisted of 150 items divided into two forms, A and B, of 75 items each. In the selection of these 150 items, item analysis data for internal consistency and difficulty of the preliminary forms were utilized.

In constructing the two forms a strong endeavor was made to balance them with regard to both difficulty and subject matter content, in order to obtain two equivalent forms. Four additional sample problems were included in both forms, making a total of eight items in each of the fore-exercises. This was done in order to minimize possible practice effect. In preceding editions of the test, a liberal time limit was allowed permitting almost all students to finish the test. No time limit was imposed on Minnesota Psycho-Analogies, in order to avoid even more stringently whatever negative influences the time factor may have on the scores of an analogy-type test. The responses to either form are indicated on a standard IBM answer sheet which may be machine scored.

One limitation should be pointed out: the population sampled for Psycho-Analogies differed in certain characteristics from the population for which Minnesota Psycho-Analogies will be used. Practical considerations determining the number and availability of subjects made it impossible to obtain the most appropriate group for the purpose of initial selection of items. However, the item difficulty was somewhat higher than 50 per cent, if appropriate correction is made for chance

success. Since the instrument would eventually be used mostly on graduate students, who, as a group, are superior to the original sample, it was considered desirable to have the average item above the 50 per cent difficulty level.

In administering these forms the major concern was to obtain adequate normative data for graduating seniors and various levels of graduate students. A secondary objective was to obtain additional course prediction data wherever possible.

Both forms of Minnesota Psycho-Analogies were administered to 33 graduating seniors and 125 graduate students majoring in psychology at the University of Minnesota. The graduating senior sample consisted of volunteers and represents only 25% of the total available population, whereas the graduate students were required to take the test and 93% of them complied. There is some possibility, therefore, that the graduating seniors tested represented a positively biased sample from their population. Half of the total group took Form A first and the other half took Form B first, thereby producing a balanced experimental design to permit estimation of whatever practice effect may accrue from taking one form first. In addition to obtaining these normative data, Form A was administered to one section of Vocational and Occupational Psychology and a section of Introductory Laboratory Psychology in order to obtain additional course prediction data.

Table 1 presents the means and standard deviations of all groups who took Form A in the spring and summer of 1949. At this time only 50 of the graduate students had been tested. Almost the entire range from a chance score of 17.5 to the maximum score of 75 is utilized. Despite some overlap, which should be expected, the upper part of the range

Table 1

Means and Standard Deviations of Groups Who Took Form A of Minnesota Psycho-Analogies

Group	N	M	S.D.
Graduate Student Majors	50	64.1	5.58
Graduating Senior Majors	33	51.7	7.37
Psychology 130	39	40.7	8.78
Psychology 5	23	32.9	7.84

Table 2

Means and Standard Deviations on Form A, Form B, and Forms A and B Combined, of Minnesota Psycho-Analogies for Four Groups of Students Majoring in Psychology

Group	N	Form A		Form B		Forms A and B Combined	
		M	S.D.	M	S.D.	M	S.D.
Graduating Seniors	33	51.7	7.37	50.2	7.15	101.9	13.78
First Year Graduate Students	50	57.7	6.91	57.3	6.67	114.9	12.76
Second Year Graduate Students	44	61.6	6.46	60.0	6.25	121.6	11.55
Third Year Graduate Students	31	66.6	4.25	65.7	5.16	132.2	7.93

is utilized by graduate student majors. This is followed in order by graduating senior majors, Vocational and Occupational Psychology, and Introductory Laboratory Psychology. It should be pointed out that Vocational and Occupational Psychology was a course open to juniors, seniors and graduate students who had at least nine quarter credits in psychology but who were not necessarily psychology majors; and Introductory Laboratory Psychology was an elementary sophomore

course. The graduate student sample was made up for the most part of graduate students who were more advanced on the average than the more complete sample reported in Table 2.

Results on Forms A and B separately and combined are presented in Table 2. The data for the total sample of 125 graduate students appear in this table. These graduate majors are broken down into first, second and third year groups depending on how much graduate work they had taken previously. The data

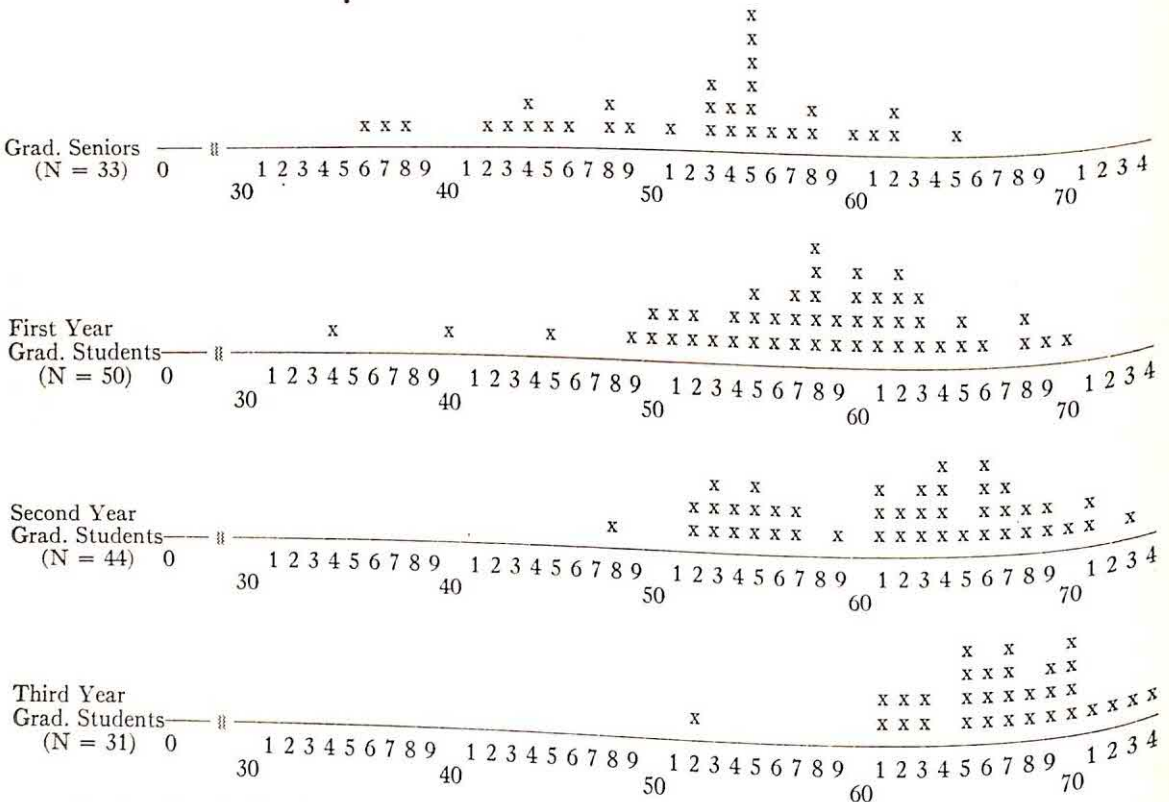


FIG. 1. Distribution of scores on Form A of Minnesota Psycho-Analogies for four groups of students majoring in psychology. Each x = one student.

indicate that the variability of scores decreases as central tendency increases in successive year groups; and that perhaps Minnesota Psycho-Analogies is most useful at the graduating senior and first year graduate student levels, and least useful for advanced Ph.D. candidates. These general trends are illustrated in Figure 1, which presents the distribution of scores on Form A. The critical ratios of the differences between means of successive groups were computed for Forms A and B combined. All of these critical ratios were significant beyond the 1% level, indicating that statistically significant discriminations are made between successive levels of psychology majors.

Norms were developed for the four categories of psychology majors for each form separately and for both forms combined. These norms are presented in the manual designed to accompany Minnesota Psycho-Analogies. Since courses of instruction and selection standards vary at different institutions, norms should be developed for each institution separately before using Minnesota Psycho-Analogies as a service instrument.

Available data tend largely to support the assumption that Forms A and B are equivalent for samples of graduating seniors and graduate students. However, there is some suggestion that Form B is slightly more difficult, but the difference is only one raw score point between means. Since these forms were constructed largely on the basis of item analysis data of a sample selected from a more heterogeneous population, equivalence will eventually have to be determined separately for different populations. At any rate, the possible difference in difficulty even for the particular populations represented by the samples studied is of a rather small magnitude from a practical point of view.

Correlation between Forms A and B for the original sample of graduating seniors and graduate students was .89. For the total sample of graduate students alone the reliability estimate dropped to .78. Using the Spearman-Brown formula, the reliability estimate for the total test of 150 items for graduate students alone is .88. There is some indication in the above coefficients that for a restricted range of talent such as graduate students,

both forms of the test should be combined in order to provide a stable enough score for individual prediction or diagnosis. More practical indicators of the stability of individual scores are obtained from the standard errors of measurement which are as follows: Form A=3.3, Form B=3.2, Forms A and B combined =4.6.

Since the difference in means between the form administered first and the form administered second was not significant even at the 5% level, it may be concluded that there is no practice effect from taking one form first when there is no time limit for taking either form.

The rather appreciable correlation of .72 was obtained between Miller Analogies, Form G, and the complete Minnesota Psycho-Analogies for the 96 graduate students on whom these test data were available. The magnitude of this correlation is partially attributable to the fact that the more advanced graduate students, because of the nature of the selection involved, tend as a group to be more "Miller-bright" as well as having more training in psychology.

For the same sample of 96 graduate students, the mean raw score of 78.9 on Miller Analogies is about as relatively high in terms of the total possible score of 100 as the mean raw score of 121.1 is on the total 150 items of Minnesota Psycho-Analogies. This comparison is defensible since both tests have four alternatives and the number right constitutes the score in both cases. This fact is pointed out since Miller Analogies, Form G, was developed especially for graduate students, and insofar as the principal use of Minnesota Psycho-Analogies will also be for graduate students, one of the main considerations is to have adequate ceiling for the upper ranges. Because items were selected for Minnesota Psycho-Analogies on the basis of an analysis of a lower mean level sample, there is some question as to adequacy of ceiling at the upper levels. It is interesting to note that the mean Miller score of 78.9 is almost at the 80th centile of graduate students in general. So it is also conceivable that the group of graduate students in the University of Minnesota's Department of Psychology is somewhat above the mean of other graduate departments in Miller ability as well as in the kinds of special training which

Minnesota Psycho-Analogies is measuring, and that therefore the norms developed here may be rather high for graduate departments in general.

Also significant is the comparison of standard deviations on both tests for the sample being considered. The standard deviation for Miller G is 8.98 and for the Psycho-Analogies, Forms A and B combined, is 12.59. Again, in view of the relative number of items, both tests may be regarded as having approximately equal variabilities for graduate students in psychology at the University of Minnesota. It is to be expected then that in those other departments of psychology where the variability on the Miller is greater, the spread of Minnesota Psycho-Analogies scores will also be greater—particularly in view of the possibility that the special informational content of Minnesota Psycho-Analogies may be unduly influenced by the subject matter content of courses in the Department at the University of Minnesota.

Also similar are the shapes of the distributions of the two tests. Both Miller and Minnesota Psycho-Analogies test scores are negatively skewed. The form of these distributions may reflect to some extent the kind of selection that has taken place, and it is a plausible hypothesis that if graduate students in psychology at the University of Minnesota were less rigorously selected, then both distributions would assume a more normal form.

A correlation of .69 was obtained between combined examination scores for Introductory Laboratory Psychology and scores on Form A of Minnesota Psycho-Analogies. For Psychology 130 an r of only .30 was obtained. The latter r is significantly lower than hitherto obtained on similar samples with preceding forms of the test. The lower r may be partially attributed to the more restricted range of the criterion, since only a two-hour final examination was given instead of the usual one-hour mid-quarter and two-hour final. Upon querying the instructor for this course, another more interesting explanation emerged. It seems that in selecting items for the final examination, an attempt was made to weed out items which appeared to be saturated with either general ability or previous background, since less time was available for testing than is ordinarily the case.

Conclusions

It may be concluded on the basis of the data obtained in this project that a special analogies test is a useful predictor of achievement in psychology courses, particularly at the more advanced levels. That the special analogies test also functions as a good terminal evaluation instrument is indicated by the rise in mean scores with increased amounts of course work and with higher levels of attainment in psychology.

As a selection instrument, Minnesota Psycho-Analogies may be conceived of as a supplement to Miller Analogies. Thus Miller Analogies may be used as one of the criteria for admission to graduate work, and Minnesota Psycho-Analogies may be employed to further determine competency to undertake advanced work in psychology. As such, both instruments will be used as successive hurdles, and the score on Minnesota Psycho-Analogies may be better interpreted in the light of Miller Analogies ability. Whether or not any student who attains a high score on Miller Analogies but who gets a low score on the special analogies should be excluded from a particular department would be a function of the available facilities at the time and the nature of the other relevant data on the individual. Theoretically, a student who gets a high score on Minnesota Psycho-Analogies should also do well on Miller Analogies. If this is not the case, it may be due to the defects inherent in analogy tests under timed conditions for certain kinds of individuals, and it should provide the basis for re-administration of Miller Analogies, as is sometimes done. Minnesota Psycho-Analogies has incorporated certain administrative advantages based on experience with the Miller. Thus, for example, there are two forms of the instrument which make certain kinds of cheating more difficult. Also, when it is necessary to re-administer the test another form may be given, thereby obviating specific memory factors. The untimed nature of the test tends also to reduce some of the sources of contamination.

The basic rationale underlying the psychological analogies test may be extended to other fields where a similar need exists. Thus it

would be possible to develop similar instruments for biology, sociology, political science, economics, and so on. A still more useful application in terms of the magnitude of the selection task would be in the realm of medical aptitude tests. A special analogies test designed to measure achievement in biology, chemistry, and physics, as well as general

ability may be developed for medical school selection purposes. Or it may be found more feasible to utilize such a test as a subtest in a more diversified instrument. At any rate, the data obtained in this project would tend to indicate the feasibility of exploring the possible uses of special analogies tests in other fields.

Received December 20, 1949.

A Note on Norms for the Purdue Industrial Mathematics Test and the Adaptability Test

Howard E. Page

CNA Tra Staff, NAS, Pensacola

The Management Engineering Division of the Naval Air Station, Pensacola, Florida is faced, from time to time, with the task of selecting from a large number of aviation tradesmen a few individuals for upgrading into positions of Planner and Estimator and Shop Planner. In the past, such selection has been made in terms of past experience in one of the aviation trades and successful and progressive experience at the Journeyman level.

Recently, personnel responsible for such selection have become interested in the use of psychological tests as an aid in such selection. Since no professionally trained psychologist is available on their staff, the writer has served in a consulting capacity on several occasions.

A major difficulty has been the non-availability of adequately standardized trade tests with normative data applicable to the aviation trades. Lack of personnel and time has precluded the development of "tailor made" tests with the result that commercially available tests have been used and norms established for the local population.

A recent administration of two tests—the Purdue Industrial Mathematics Test and the

Adaptability Test by Tiffin and Lawshe—provided data on a population of 152 Aviation Tradesmen with a sufficiently large representation from two trades to make feasible the publication of such norms.

Table 1 summarizes the background data on this population. It is to be noted that 48 per cent of the total population have served an apprenticeship in a trade, and that 71 per cent have completed Trade School training. Fifty-seven per cent of the population has completed at least a high school education. This is a larger percentage than might have been expected. On the other hand, the group is a relatively young group in terms of age and in number of years on the job. The range for years on the job was from two months to 22 years. It must be concluded that personnel included in this group are well trained and experienced in their particular aviation trade.

Table 2 presents the mean scores and standard deviations for those trades where N was large enough to make this meaningful. ASTP trainees do somewhat better on the Industrial Mathematics test and the Vocational Trade School students do less well than does the

Table 1
Descriptive Data for Population Tested

	N	No. Who Served Apprentice- ship	No. Who Had Trade Sch. Trng.	No. Having at Least HS Educa.	Average No. Yrs. on Job	Average Age
Aviation Mech. Gen.	51	9	34	32	5	34
Metalsmith	42	21	32	26	6	35
Machinist	17	15	10	5	7	37
Electrician	16	13	11	9	6	38
Aircraft Engine Mech.	12	7	11	6	5	37
Instrument Mechanic	6	1	4	5	4	33
Radio Mechanic	3	1	3	3	9	39
Electroplater	3	0	1	0	6	43
Unclassified	2	1	2	0	2	29
Total	152	68	108	86		

Table 2
Mean Grade and Standard Deviation by Trade

Trade	N	Adaptability Test		Industrial Math Test	
		Mean	S.D.	Mean	S.D.
Aviation Mech. Gen.	51	16.8	5.9	17.4	5.7
Metalsmith	42	15.9	4.6	17.0	5.5
Machinist	17	16.5	7.2	15.6	6.6
Electrician	16	15.3	5.4	15.3	5.7
Aircraft Eng. Mech.	12	14.2	3.6	13.7	3.9
Total	152	16.1	5.6	16.4	5.6

population reported. This is as one would predict in terms of the higher educational level of the ASTP students as compared with the Aviation Tradesmen and the latter's greater "on the job" experience as compared with Vocational Trade School students. On the Adaptability Test the mean scores for the Aviation trades are comparable to those published by the authors for Naval Electrical Trainees, Employees of a Piston Ring Manu-

facturing Company and Total Group. Aviation Tradesmen surpass Female Applicants and fail to do as well as Employed Clerical Workers and Purdue Seniors.

Table 3 presents the norms for the Adaptability Test. Data to the left of the vertical line are reproduced from *Examiners Manual for the Adaptability Test* by Joseph Tiffin and C. H. Lawshe and published by Science Research Associates, 228 South Wabash Avenue,

Table 3
Norms on the Adaptability Test

Percentile	Standard Score	Female Applicants	Navy Electrical Trainees	Employed Clerical Workers	Purdue Seniors	Employees of a Piston Ring Mfg. Company	All Groups Combined	Metalsmith	Avia. Mech. Gen.	Total Aviation Tradesmen
100	3.00	27	32	35		33	34	30	35	33
99	2.32	23	29	32	35	30	30	27	31	29
98	2.05	22	27	30	34	29	28	26	29	28
95	1.64	20	25	28	32	26	26	25	27	25
90	1.28	18	24	26	31	23	23	22	24	23
80	.84	15	22	24	29	20	20	20	22	21
70	.52	13	20	22	28	18	18	18	20	19
60	.25	12	19	21	27	16	16	17	18	18
50	.00	11	17	20	26	14	15	16	17	16
40	-.25	9	16	18	25	11	12	15	15	15
30	-.52	8	15	17	24	9	10	13	14	13
20	-.84	6	13	15	23	7	8	12	12	11
10	-1.28	4	11	13	21	4	5	10	9	9
5	-1.64	2	9	11	20	3	4	8	5	7
2	-2.05		7	10	18	1	1	6	4	5
1	-2.32		6	8	17			5	3	3
Total		N = 612	640	86	43	794	2175	42	51	152

Table 4
Norms on Purdue Industrial Mathematics Test (Form A or B)

Percentile Score	Std. Score	ASTP Trainees	Vocational Trade Sch. Students	Metal-smith	Aviation Mechanic General	Total Aviation Tradesmen
100	3.00	32	28	33	35	33
99	2.32	31	27	30	31	30
98	2.05	30	26	28	29	28
95	1.64	28	23	26	27	26
90	1.28	27	20	24	24	24
80	.84	25	18	22	22	21
70	.52	23	16	20	20	19
60	.25	21	14	18	19	18
50	.00	20	13	17	17	16
40	-.25	18	12	16	16	15
30	-.52	17	10	14	14	13
20	-.84	16	9	12	13	12
10	-1.28	14	7	10	10	9
5	-1.64	13	6	8	8	7
2	-2.05	11	5	6	6	5
1	-2.32	10	4	4	4	3
Total		N = 125	188	42	51	152

Chicago, Illinois. To the right of the vertical line are comparable norms for Metalsmith, Aviation Mechanic General and the total Aviation Tradesmen tested at Pensacola.

Table 4 presents the norms for the Purdue Industrial Mathematics Test (Form A or B). Data to the left of the vertical line are reproduced from *Preliminary Manual—The Purdue Industrial Mathematics Test* by C. H. Lawshe, Jr. and Dennis H. Price—and distributed by the Division of Applied Psychology, Purdue University, Lafayette, Indiana. To the right of the vertical line are the additional norms developed at Pensacola.

The correlation between the Adaptability Test and the Industrial Mathematics Test for the total population ($N=152$) proved to be .71. This seems very high for the two tests, where one is supposedly measuring arithmetic ability while the other attempts to measure general aptitude. An inspection of the items in the two tests, however, shows considerable overlap which would account for the high relationship found.

The reliability of the two tests calculated by the Kuder-Richardson short formula was found to be .74 for the Adaptability Test, and .76 for the Industrial Mathematics Test.

Received December 5, 1949.

Measurement of a Complex Psychomotor Performance by Means of a Printed Test

Lloyd S. Nesberg and Karl U. Smith

University of Wisconsin

A practical aspect of measurement of human capacities and traits is the simplification of present methods in appraisal of psychomotor performance. One possible approach to simplified design of the psychomotor test is the development of printed examinations which will scale sensory-motor capacity in the same general manner as the apparatus test. In the present study, developmental research on a printed test has been carried out in order to measure the performance involved in a judgmental reaction time test, the Vector Complex Reactometer. This test has been developed for determination of some aspects of pilot aptitude.

Methods

1. *The Vector Complex Reactometer.* In this test (Figure 1) the subject's reaction time is measured in turning a series of switches relative to the changing position and direction of three lights, the pattern of which is altered during a predefined sequence. The subject, in turning a given switch on the response board, selects a correct group of switches in terms of the relative direction of a red light with respect to a green light and, thereafter, a particular correct switch within the group in terms of the position of a white light. The three lights are presented on the stimulus panel automatically, and the subject's response immediately prepares the apparatus for the presentation of the next stimulus pattern.

The test under consideration here is designed to present forty different light patterns in a sequence, which may be repeated after the sequence is completed. It may be administered with a definite time limit, in which case the number of reactions in a given test period are scored, or the score may be defined as the time required to perform a given number of reactions. In the research conducted, the test was scored in terms of the total number of

correct switches turned in a four-minute test period.

2. *The Motor Decision Test.* In designing a printed form of this test (Figure 2) the same general principles of stimulus presentation and response involved in the Reactometer have been incorporated into each test item. Reference to the sample item in Figure 2 will show how the test was designed. Instead of the four groups of five white lights in the stimulus panel of the apparatus test, the printed item presents four groups of circles each arranged as in the performance test. The critical stimulus among these groups of white circles is indicated by the black filled circle, as shown in the upper right hand group in the sample item. In the printed item, triangles and squares are substituted for the red and green lights of the apparatus test, and a black triangle and a black square constitute the critical stimuli for guiding the response.

In the printed test, the subject responds by checking one of twenty-five inverted *T*'s. These inverted *T*'s represent the switches in the performance test, and are grouped in banks of five as in the performance test. The correct bank of switches is indicated by the relative position of the black triangle with respect to the black circle. The particular correct switch within this bank is indicated by the position of the black circle.

The Motor Decision Test reproduces all forty stimulus configurations of the apparatus test. Some of these forty items are repeated in the test to give a total of 108 items. The completed form of the test consists of 12 pages of test items, with 9 items on each page. The score on this is the number of correct reactions minus the number of items marked incorrectly.

Experimental Results

The treatment of experimental results will be discussed in the following order: (1) the

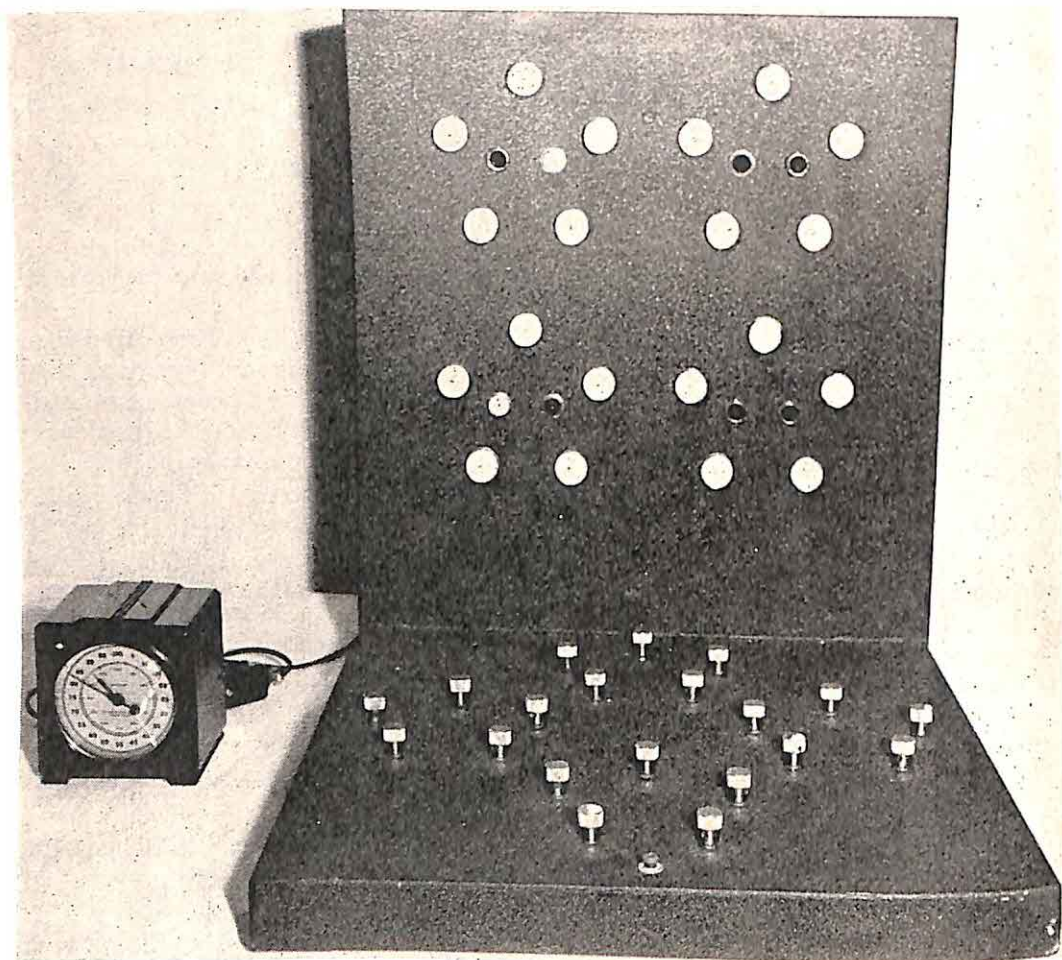


FIG. 1. The Vector Complex Reactometer. This device, designed by Dr. Jack Buel, School of Aviation Medicine, San Antonio, Texas, is manufactured by the Vector Manufacturing Co., Houston, Texas. The instrument was obtained for research through the cooperation of Dr. Buel.

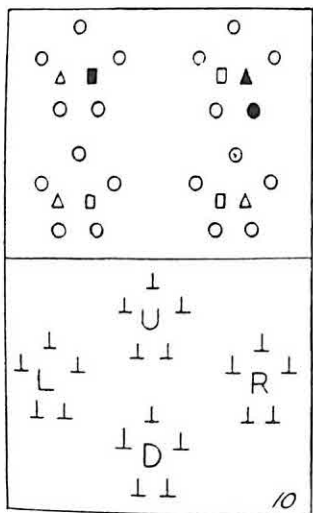


FIG. 2. Design of a sample item in the Motor Decision Test.

standardization of the Motor Decision Test; (2) correlations and interrelations of the performance and printed tests; (3) the amount and significance of transfer; and (4) sex differences in performance.

1. *The Standardization of the Motor Decision Test.* Two preliminary forms of the Motor Decision Test were designed and investigated before a final copy of the test was drawn up. On the basis of data obtained with these two forms of the test as well as with the final four-minute form, the following facts were established. The four-minute test interval established for this test length allowed none of the subjects to solve all of the stimulus patterns. Memory and fatigue effects were negligible. Analysis of the location of errors does not reveal any differences in item difficulty. Errors of

Table 1

A Quantitative Summary of Performance on the Printed and Apparatus Tests

	Sex	Range	Mean	S.D.
Group I Performance Test (50 subjects)	25 Males	27-99	75.3	19.39
	25 Females	37-99	74.3	15.47
Group II Performance Test (50 subjects)	25 Males	58-135	85.8	13.51
	25 Females	58-105	82.3	13.55
Group I Paper-Pencil Test (50 subjects)	25 Males	28-98	61.6	14.46
	25 Females	27-83	54.2	14.57
Group II Paper-Pencil Test (50 subjects)	25 Males	37-100	65.7	17.27
	25 Females	44-91	66.2	12.72

response to specific test items are not persistent. Total errors are extremely small, hence the number of items marked represents the major factor in the scoring.

Distributions of scores on both performance and printed tests are approximately normal. A detailed quantitative description of data is presented in Table 1. This table gives the means, range and standard deviations in two groups of subjects for both sexes and all test sequences. Group I was given the performance test first and, 48 hours later, the printed test. Group II was given the printed test first. The subjects were college students. There are no great differences in the range of scores on the printed and performance tests. The means are consistently lower on the paper and pencil form regardless of its position in the testing order. Males as a whole are more variable in test performance than females. There is no consistent trend in variability with regard to test sequence.

Reliability. The test-retest reliability of the Vector Complex Reactometer, computed from the scores of a separate group of 23 subjects, is +.86. This is the reliability found when a one-day period separated the two administrations. The test-retest reliability of the Motor Decision Test, based on the performance of 53 subjects and a temporal separation of 7 days, is +.83. Data presented below will show that correlation between the apparatus and printed tests approaches the reliability of each, as just described.

2. *Correlation between the Performance and Printed Tests.* Correlations between the two

tests were calculated for males and females separately and for the two conditions of test sequence used. These data are given in Table 2. All values given are significant at the 1 per cent level of confidence. When the correlation values given are transformed into Fisher's Z ,¹ and the standard error of the Z 's computed, the differences are not statistically significant. It is therefore reasonable to suggest that for the conditions given, the Motor Decision Test displays a high level of duplication of measurement of the factors involved in performance on the Reactometer.

3. *Transference of Response between the Two Tests.* The design of the experiment permitted determination of the degree to which prior performance on the printed test affected scores on the performance test and vice versa. This

Table 2

Correlations between the Performance and Printed Tests*

Groups	r
Composite for females	+0.70
Composite for males	+0.69
Females: performance test first	+0.82
Males: performance test first	+0.79
Females: performance test second	+0.84
Males: performance test second	+0.64
Composite: performance test first	+0.79
Composite: performance test second	+0.77

* All values significant at the 1% level of confidence.

$$^1 Z = 1.5313 \log \frac{1+r}{1-r} (8).$$

transference may be attributed to general situational learning occurring in one test situation which is carried over to the second test, or to specific skill acquired in one test situation which is applicable to the second test. The amount of transference of response or generalization is indicated by the change in performance on one test attributable specifically to the fact of prior performance on the other test. When all subjects are considered together, a statistically significant increment in performance on the printed test is found when this test is taken after the performance test. The level of confidence of the difference in scores on the performance test when taken after the printed test and when taken without such previous testing experience also exceeds the one per cent level. Analysis of variance of the transfer data confirms these general statements and indicates, furthermore, that there is no significant effect of test sequence on transfer. This analysis discloses, however, that there are certain inherent differences in the measured characteristics of the performance and printed tests.

4. *Sex Differences in Performance on the Two Tests.* Values of t for differences in performance for males and females were computed for all data on each test and also for data separately obtained on each test with respect to test sequence. None of the values was significant, which confirms the hypothesis that the effect of the sex characteristic on these two tests is zero.

Summary

A printed test, designated the Motor Decision Test, has been designed to duplicate a complex reaction time apparatus test which has been manufactured for study of aircraft-pilot aptitude. The apparatus test has been given the name of Vector Complex Reactometer.

Utilizing an item design made up of differently shaded forms which simulate light posi-

tions and switch arrangements in the apparatus test, it has been possible to duplicate in the printed test all combinations of stimuli provided in the apparatus test. Comparisons of the two tests are made in terms of range of test scores, relative reliability, interrelation of test scores, and degree of generalization of performance from one test to the other.

Results show that the two tests have approximately equal test-retest consistency. The reliability of the apparatus test is $+.86$, that for the printed test $+.83$. These values are only slightly higher than those found for the intercorrelation between the two tests. For different groups of subjects, the intercorrelation values typically vary from $+.70$ to $+.80$.

Means of scores on the printed test, expressed as total number of items answered correctly, are usually somewhat lower than those on the apparatus test. In the latter case, scores are expressed as the total number of effective responses made in a standard time. The distribution and variability of scores for the two tests are not significantly different.

An analysis has been made of the degree to which performance on one test may affect scores made on the other. A significant positive transfer effect is found for both tests in terms of the effect of the prior administration of one on the test scores found in the later performance on the other. Some significant minor differences in the measured characteristics of each test are found through such overall analysis of the tests when administered successively.

In general, it has been concluded that the printed Motor Decision Test duplicates extensively results obtained with the complicated apparatus test, and that such a printed test could be used safely as a substitute for or in conjunction with the apparatus test in screening procedures in which complex reaction time may be of practical significance.

Received January 6, 1950.

Visual Skill and Performance in a Meat Packing Plant *

F. Nowell Jones

University of California at Los Angeles

and

Charlotte Jean Smith

United States Spring and Bumper Company

To the best of the writers' knowledge, no work on the testing of packing house employees has ever been reported. It would seem that there would be considerable possibility for the use of selection devices in this industry, since many of the jobs require a high degree of skill, and involve some hazard of injury, especially cuts.

Preliminary Study

The work on visual skill grew out of a preliminary study of a casings inspection department. Here the job required size grading and inspection by visual control, and results were negative when production was compared with the various Ortho-Rater measures, despite the fact that the criterion of production, output for two successive weeks, had a reliability of .96.¹ However, N was only 17, the total number of workers in the department, and so it seemed desirable to extend the study to other jobs in the plant.

The Present Study

The Jobs. The three jobs selected for study were wiener skinning, bacon slice and wrap, and shipping cooler work. Wiener skinning is the procedure of removing the cellophane

"casings" from "skinless" frankfurters or wieners after smoking. This involves the use of a moderately sharp knife to cut the links apart, and to start the tearing of the cellophane. The skin is then removed by a spiral tearing action. Study of this job had been completed by the time study department, and there was considerable interest in improving production. The cellophane was quite difficult to see, being transparent, and it is possible that the job is performed best by "feel."

In the bacon slice and wrap operation we were concerned with the persons who lifted sliced bacon from a conveyer belt, weighed it into pound lots, and wrapped each lot in paper. Each operation involved visual control.

The shipping cooler jobs were of the laboring type, in that the main work was moving stock to fill orders. It would not be expected that close visual control would be necessary here.

Criteria. In the wiener skinning department it was possible to obtain two independent ratings of each employee, one by the foreman and one by the time study man who had been regularly assigned to this department. They rated each worker on a percentage scale, with 100 indicated as the norm. Each rater made two ratings, a month apart. Unfortunately, the second rating by the foreman disappeared in the mails, and was therefore irretrievably lost. The two ratings by the time study man correlated .94, and the average of his two ratings correlated .81 with the foreman's rating.

In the bacon slice and wrap department the criterion consisted of the foreman's rating on a 5-point scale. As a matter of fact, the 1 position was not used, on the ground that all such workers had been transferred out. A second rating was made by the same foreman some weeks later, and the two ratings correlated .84. Considering the rather short scale used, this indicates a reasonably high criterion reliability.

* This article is based on part of the material submitted by the junior author in partial fulfillment of the requirements for the M.A. degree at the University of Wisconsin. We wish to thank the Bausch and Lomb Optical Company for making available the Ortho-Rater used in this study, and to thank Mr. Harold Jaeke, superintendent, and Dr. John M. McGinnes, psychologist, of Oscar Mayer and Company, Madison, Wisconsin, for respectively permitting us to use the facilities of the plant, and assisting us in obtaining subjects and criterion scores.

¹ The Wonderlic Personnel Test and the Minnesota Rate of Manipulation Test were also administered to this group. The correlations with the criterion were: Wonderlic, -.31; Minnesota, placing, -.30; Minnesota, turning, -.39. These tests are apparently worthy of further study.

Table 1
Correlations Between Production and Criterion

Test	Correlations*		
	Wiener Skinners	Shipping Cooler Employees	Bacon Slicers and Wrappers
Ortho-Rater:			
Far Vertical Phoria	.32	-.13	.09
Far Lateral Phoria	.31	-.12	-.29
Near Vertical Phoria	.02	-.06	.08
Near Lateral Phoria	-.03	.06	-.03
Far Acuity, Both Eyes	.37	.09	-.09
Far Acuity, Worse Eye	-.19	.24	.00
Near Acuity, Both Eyes	.05	.19	.07
Near Acuity, Worse Eye	-.18	.23	.13
Depth Perception	.20	.03	.03
Color Perception	-.10	-.02	.04

* None of these correlations is significant; in no case does r/σ_r approach 2.58, and t 's show insignificant r 's for wiener skinners.

The criterion in the shipping cooler was not at all satisfactory. The foreman's rating was used as in the case of bacon slice and wrap, but only the 3 middle positions on the scale were used, and the correlation between the first rating and a rerating was only .45.

Subjects. In bacon slice and wrap and wiener skinning the employees were almost entirely female. A further restriction in bacon slice and wrap was the hiring of only young workers. In the shipping cooler, the employees were predominantly male. The populations were as follows: bacon slice and

wrap, 47; wiener skinning, 26; shipping cooler, 66.

Testing. Testing with the Ortho-Rater was carried out in each department. Taking the test was voluntary, but as a matter of fact there were no "holdouts." If glasses were customarily worn on the job, they were worn for the test. Testing was on company time.

Results

Table 1 shows the correlation between the criterion in each department and the various scales on the Ortho-Rater. None is signifi-

Table 2

A Comparison of the Means and Standard Deviations for the Packing House Group and the OSRD Group

	Packing House		OSRD	
	N = 139		N = 234	
	Mean	σ	Mean	σ
Far Vertical Phoria	5.43	1.31	5.44	0.85
Far Lateral Phoria	7.59	3.14	7.19	1.97
Far Acuity, Both Eyes	9.45	2.49	10.96	1.37
Far Acuity, Worse Eye	8.15	2.46	9.58	1.68
Depth Perception	2.74	2.95	4.39	3.00
Color Perception	4.30	1.06	4.81	1.01
Near Acuity, Both Eyes	9.50	1.67	11.91	1.10
Near Acuity, Worse Eye	8.50	2.23	10.50	1.64
Near Vertical Phoria	4.14	1.55	4.30	1.05
Near Lateral Phoria	8.57	3.86	7.97	2.25

cant at the 5% level, when tested by the standard error. The correlations for wiener skinning were also subjected to a *t* test, and were found not to be reliable. Since a correlation between a phoria test and a criterion would be hard to interpret if the phoria scores were taken straight through from high to low, the *r*'s reported here are based on scores as deviations from the norm. Correlations calculated straight through show no significant relationship either. In addition, each scatter diagram was carefully inspected, and in some cases, where it appeared useful, average performance for each score on a given Ortho-Rater scale was plotted, to determine whether or not the correlation had failed to reveal a cut-off point, or other relationship. In no case was this true.

It is apparent that, within the limits of criterion validity and reliability, application of the Ortho-Rater would, on the face of the matter, add nothing to the selection of workers for the jobs under consideration. It is possible that some restriction of range of Ortho-Rater scores might have lowered the relationships found and so we have compared our means and standard deviations with those reported

in the OSRD study of reliability.² The means are quite comparable, and our standard deviations are, except for depth perception, larger. The means and standard deviations are given in Table 2. It is difficult to escape the conclusion that the Ortho-Rater would not be useful in this company, at least as far as the jobs under consideration are concerned.

Summary

1. The Ortho-Rater was used to measure visual efficiency of employees in three packing house departments: wiener skinning, bacon slice and wrap, and the shipping cooler. In all, 139 employees were tested.

2. No significant correlation between Ortho-Rater scores and efficiency, as determined by foremen's ratings (and in the wiener skinning department, by foreman's and time study man's ratings), was found.

Received November 7, 1949.

² Adams, D. K., Beir, D. C., Imus, H. A. *A test-retest reliability study of the Bausch and Lomb Ortho-Rater with Naval personnel.* Office of Scientific Research and Development, No. 3969, August 1, 1944.

The Myth of Chronological Age *

Austin S. Edwards

University of Georgia

In recent years there has been increasing study of the problems of age and aging (3, 4, 5, 7). Increasing numbers of the aged appear in our population and old age counseling has become a definite kind of psychological work. Many articles have appeared, but except for occasional reference, no accurate information is given concerning senescent who are reasonably healthy as compared with those who are seriously deteriorated or disabled (2). Generally no distinction is made. See, for example, the excellent article on chronological age and quality of literary output (5).

This paper is concerned with the question of the differences between senescence and senility. It suggests that here, as elsewhere, chronological age is an inaccurate indication of ability and competence. To what extent do the senescent differ from the seniles? Are many senescent still quite as capable as so-called younger individuals? In what respects are healthy and relatively uninjured senescent as capable or more so than younger people? All of these questions have had insufficient answers.

It is the purpose of this paper to consider one very important aspect of ability, namely, body, hand, and arm steadiness and to give some accurate information as to the difference in these respects between senescent who are not known to have any serious physical handicaps, disease history, etc., and seniles, so diagnosed at a state mental hospital, and to compare both with the average for younger people.

Body Sway. A limited number of Ss were measured for body sway on two different occasions with no special reference to disease history. In both cases the senescent individuals were practically as steady as the younger Ss, whose ages ranged from 10 to 29. In the first study, which gave age differences, the Ss aged 50-69, twenty-four in number, showed

no greater body sway than did the younger Ss the mean and the median giving contradictory results. In the later study with 23 Ss, aged 50-70, the difference, if any, was in favor of the older Ss. This was true with eyes open; the results with eyes closed showed little difference, although some of the older Ss showed somewhat more sway than the younger.

With insufficient numbers of cases, it does not appear that Ss from age 50 to 70, who have not suffered from disease or accident to any great extent, have necessarily more body sway than do younger Ss. In fact, taking individual cases, many of the older Ss have considerably less body sway than do many of the younger.

Hand and Arm Steadiness. Finger Movements. A study has been made in connection with finger movements. The writer's finger tremometer was used with standard procedure (7). The tremometer measures finger movements in three dimensions of space. Time of measurement is 30 seconds. It is believed that finger tremor (gross finger movements) is a decidedly sensitive indicator. The mean finger tremor of 1000 Ss (aged 16-35) is 35.3 mm., S.D. 20.25. Examination of 65 senescent, aged 60-85 (average chronological age, 70.4), who had no serious amounts of disease history, and who may be considered as reasonably healthy individuals, had a mean finger tremor of 42 mm., S.D. 24.5. The difference between the means of these and of the 1000 younger subjects showed an increase of finger movements for the older group of only 19 per cent, with a critical ratio of 2.21 (significant at 5 per cent level). Some of the oldest Ss were actually the steadiest.

In contrast to this are the results of the measurements of the senile group at a state mental hospital. These 89 cases were diagnosed in the case histories by the hospital staff as senile. Their ages were 54-90 (average chronological age, 67.7). Although the average chronological age of the seniles was almost three years less than the average of the senes-

* Thanks are due Miss Anne Gilbert for assistance in part of this study.

cents, the average finger tremor was 133.3 mm., S.D. 66.8, which is more than three times as great as that of the former. The difference between the means of the seniles and the 1000 younger Ss gave a critical ratio of 12.8, and that between the senescents and the seniles, a critical ratio of 12.0.

Thus we find a small but statistically reliable difference between the senescents and the younger Ss, although the senescents cannot be said to have abnormal finger tremor or an amount that might be expected to have much, if any, practical significance. On the other hand, the seniles form a distinctly different group as compared with either of the others, and have an amount of finger and arm movement (involuntary, uncontrolled movement) that would be expected to have serious consequences so far as skilled work involving fine muscular control and steadiness is concerned.

Men vs. Women. The differences between men and women are not remarkable except perhaps on one point. The senile men had increased finger tremor in comparison with women in almost the same relative amount as occurs with our so-called younger normal groups. On the other hand, the senescent women showed relatively more increase than did the senescent men. The average for younger men is 39.83, S.D. 21.17, and for the senescent men the average is 42.67, S.D. 26.2. The norm for younger women is 30.33, S.D. 18.02, and the average for the senescent women was disproportionately higher, namely, 41.27, S.D. 26.4. If this difference between senescent men and women is truly indicative of other possible sex differences for the aged, it may be of no little scientific and practical significance.

Discussion

The competence of so-called aged people, whatever that may mean, is to be understood as an individual matter. There are many individuals above 45 who are not competent to do good skilled work. On the other hand, there are skilled workers of much greater age who are entirely competent and perhaps better both in quality and quantity of production than many younger workers. The personnel problem is not one of merely grouping all indi-

viduals above some certain age and calling them incompetent. Our study gives emphasis to the need for discovering what old people are entirely competent, and what aged people have been made incompetent not because of chronological age but because of sickness, injury, or deterioration. It is reasonable to expect that many older people need more frequent rest periods than do younger workers. It is even more important to recognize the fact that many older workers are of the highest competence and are often found to be superior in judgment, foresight, carefulness, freedom from accidents, and trustworthiness. For our study old age has meant 60-85, with practically as good hand steadiness as that of students aged 16-35. What does old age mean as it is commonly used in business and industry? Is it much more than a superstition held by those who still continue in the fossilized thinking of by-gone ages?

Summary

1. So far as we have data on body sway, it appears that many people whose ages reach at least 70 are no less steady standing in the erect position than are many younger people. A considerable number of senescents are considerably more steady than the younger Ss.

2. In our experiments upon finger tremor it is clear that senescents do not differ greatly from the average of 1000 college students. It is clearly evident that the average finger tremor for senile patients is more than three times as great as that of presumably healthy senescents whose average age was a little greater than that of the seniles (70.4 as compared with 67.7).

3. Senescent women were found to have disproportionately greater increase of finger tremor than senescent men. Although this was not great, it suggests a problem worthy of further research to discover whether greater deterioration and inability appear in aging women than in aging men.

4. It is suggested that competence for various kinds of work can only very inadequately be judged in terms of any such rough and inaccurate indication as that given by chronological age; and that abilities of men and women should be decided by means of such

accurate methods as are here suggested in one area of human behavior.

Received November 10, 1949.

References

1. Edwards, A. S. The finger tromometer. *Amer. J. Psychol.*, 1946, **59**, 273-283.
2. *Geriatrics*. Vols. 1-4, 1946-49.
3. Kaplan, O. J. (Editor). *Mental disorders in later life*. Stanford University: Stanford University Press, 1945.
4. Kuhlen, R. G. Age differences in personality during adult years. *Psychol. Bull.*, 1945, **42**, 333-358.
5. Lehman, H. C., and Heidler, J. B. Chronological age vs. quality and literary output. *Amer. J. Psychol.*, 1949, **62**, 75-89.
6. Pressey, S. L., Janney, J. E., and Kuhlen, R. G. *Life: A psychological survey*. New York: Harper & Bros., 1939.
7. Ruch, F. L. Adult learning. *Psychol. Bull.*, 1933 **30**, 387-414.

Reading Ease of Commonly Used Tests *

Ralph H. Johnson

Veterans Administration, Minneapolis

and

Guy L. Bond

University of Minnesota

The authors of this paper are concerned with the use of vocational tests requiring reading skills in the expanding post war counseling programs in personnel selection, school guidance programs, and the vocational counseling of veterans.

Numerous intelligence tests requiring reading skills are being administered to general population groups, veteran and non-veteran, and similar tests continue to be administered to students at all grade levels in survey testing for individual guidance purposes. Recent developments in the area of vocational counseling have also been characterized by the increased use of the Kuder Occupational Preference Record at the junior high school level and by more frequent use of the Strong Vocational Interest Blank at the senior high school level, plus the extensive use of both tests in the vocational counseling of World War II veterans. It appears that some clarification is needed concerning the reading levels of the general population and student groups and the extent to which vocational tests which require reading skills match these reading levels.

This article approaches the problem by indicating the general types and prevalence of reading limitations among the academic and general population. Second, this article attempts to arrive at a relative determination of the readability level of tests commonly used in counseling and group testing situations. Third, this article suggests observations and conclusions that may have implications of importance concerning the use of tests requiring reading abilities in vocational counseling situations.

* The opinions expressed herein are the views of the authors and are not to be construed as representing the Veterans Administration.

Clients assigned to tests requiring the use of reading skills may be limited in their comprehension of written materials because their intellectual capacity prevents them from comprehending material written beyond the ninth and tenth grade levels; or, the clients may have reading disabilities which precluded their reading growth from developing at the same pace as their intellectual development. A client in the latter category is considered a reading case if there is a significant degree of difference between his mental age and his reading age. At the junior high school level, a student is considered to be a reading case if his reading age is two or more years below his mental age (20). Mental age in the above definition is considered to be the mental age derived from an individual test of mental ability such as the Stanford-Binet or Wechsler-Bellevue.

Table 1
Reading Level of Adults with IQ's Below 100

IQ	MA	Reading Grades
75	11.3	5.4
80	12.0	6.2
85	12.9	7.0
90	13.6	7.7
95	14.4	8.6
100	15.0	9.2

In the general population, the prevalence of reading limitation ascribed to lack of intellectual capacity is reflected by Table 1. Table 1 is derived from Gates' (13) table for translating age scores into grade scores. Bond (5) states that normal reading grade at maturity, as indicated by Table 1, may be considerably increased by exposure to good teaching.

Other indications of the reading level of the general population are the following: (a) Unpublished results of Army and Navy surveys which indicate mean reading grade levels of between grades 8 and 10 for veterans of World War II; (b) Lorge and Blau (19) found an average reading grade level of 9.2 among 242 tested adult WPA subjects; (c) Census data for 1940 give grade 8.4 as the average school grade completed by our population. This is of significance when considered with Witty's (24) conclusion that the average attainment in reading of elementary school graduates harmonizes closely with grade expectancy.

An indication of the prevalence of reading disability cases is reported in a study by Gray (14), who states that one out of five junior high school students in Chicago had a reading disability or was considered a reading case. Another study by Monroe (20) indicates that four out of five reading cases are boys.

Assuming that these studies generally represent an approximate indication of the reading limitations of the general population, it is incumbent upon counselors to exercise care in the selection and interpretation of tests which require varying degrees of reading skills. Since test results may vary depending upon the reading level of the client, estimating and, if possible, knowing the approximate reading level of the client and the readability level of the test seems necessary to insure reasonably accurate test selection and interpretation.

Clues which may give a general indication of the client's reading level are the following: (a) Results of a test of mental ability, if available; (b) Vocabulary level of the client as revealed in the interview; (c) Client's facility at interpreting questionnaire material; (d) Content of client's correspondence; (e) Client's educational level (use with caution, since an individual's reading level may vary as much as six grades from his educational level).

Observation by the psychometrist or counselor of the client's behavior in the test room situation may prove very meaningful in detecting clues as to the reading level and type of reading disability characterizing the client. These clues may consist of the following: (a) Excessive articulation and lip reading; (b) Use of crutches in reading (pointing with finger and pencil); (c) Frequent regression in reading

passage; (d) Excessive fixations per line; (e) Frequent consultation with psychometrist regarding comprehension of test directions and items; (f) Speed of reading; (g) Discrepancies between verbal and non-verbal tests and discrepancies between subtests; (h) Complaints of tired eyes.

A quick check of the difficulty of a passage for an individual is the number of words he misunderstands and misses orally out of 20. Betts (4) states that if a student in the elementary or secondary grade level misunderstands over one word in 20, he may lose the meaning of the passage. This type of check may prove efficient in a testing situation. If there is sufficient time, group or individual reading survey tests would give the desired information.

Assuming the counselor's selection of individual tests in a test battery is made to conform with the client's reading level, one must make the further assumption that the counselors have information regarding the reading level of various vocational tests.

The authors have felt that more information was needed regarding the reading level of various tests which are being given rather routinely to academic and general population groups, and have attempted to measure their reading level by applying the Flesch formula to the more commonly used tests in vocational counseling (2, 3). In most instances, the formula was applied to the directions separately.

Formulas for measuring the readability level (comprehension) of grade school text books have been in use for the past 25 years and Klare (17) states that to date there are about 34 formulas or methods available.

The Flesch formula was used in this limited survey for the following reasons: It is an efficient formula; its author (9, 11) claims it has been used with success in sampling the reading level of adult reading materials; it has been applied experimentally to the readability level of Public Opinion Questionnaires (16), and Klare (17) states that it correlates significantly with other formulas. Flesch's revised formula (11) is based on the number of syllables per 100 words and the average sentence length in words. Scores resulting from this formula vary from 100 to zero. A score of 100 corre-

sponds to the prediction that a child who has completed the fourth grade will be able to answer $\frac{3}{4}$ of test questions asked about a passage that is being rated. A score of zero indicates that the passage is practically unreadable.

For further information regarding development, standardization, reliability, and validity of this formula, the reader is referred to publications by Flesch (9, 10, 11, 12).

The application of the formula to some commonly used vocational tests gave the results indicated by Tables 2 and 3. Table 2 represents tests for which an overall selected sample was used in determining readability level. Due to the spiral nature of the tests listed in Table 3, three separate measures of reading ease were made.

In determining the Reading Ease of multiple choice test items, only the correct response was counted. Had all five of the possible responses been included, sentence length and word length would have been increased, resulting in a lower (more difficult) Reading Ease score. Therefore, the results listed in Tables 2 and 3 tend to represent a conservative

estimate of the reading difficulty of the test items sampled.

The authors recognize that the application of the Flesch formula to test items in vocational counseling test batteries was not the purpose for which the formula was designed. Therefore, the following definite limitations of the formula must be considered in any interpretations resulting from its application: (a) Reading grade placement scores above grade 7 represent estimated corrections for area of extrapolation beyond grade 7; (b) The formula does not appear to measure the effect of complexity caused by double negatives in a test such as the Minnesota Multiphasic Personality Inventory; (c) The formula does not appear to measure the rather involved directions in some parts of the Strong Interest Test; (d) The above results have not been verified by any extended sampling and, with the exception of the interest test items, sampling is limited to test items that are complete sentences; (e) The formula is not designed to measure the readability level of test items that are not complete sentences. Therefore, the readability level of the Strong, Kuder, and California Interest Tests, as indicated in Table 2, can be considered valid only in the sense that they indicate relative difficulty of the several tests.

It appears that a formula which measures word complexity and abstractness may be more appropriate in measuring the readability level of interest test items which contain single words and phrases. The authors did not attempt to apply a formula of this nature to the interest tests. However, the Lewerenz formula, which is based on vocabulary alone, was used by Steffle (21) in computing the reading difficulty of interest inventories and, with the exception of the results on the Study of Values, there appears to be relative general agreement on the readability level of interest tests as measured by the Flesch and Lewerenz formulas. It is of interest to note that Aukerman (1) questions the importance of specific vocabulary ability in reading situations, since his study indicated that good students were significantly superior to poor students in general reading ability, and the fact that good students and poor students were not significantly different in either general or specific

Table 2

Over-all Readability Levels of Selected Tests
as Determined by Application of
the Flesch Formula

Test	Reading Ease	Grade Level
Bennett Mechanical Comprehension	90	5.5
Minnesota Multiphasic Personality Inventory	88	6.0
Directions for Minnesota Clerical	87	6.0
Bell Adjustment Inventory	80	7.0
Directions for Bell	61	9.5*
California Interest Inventory	65	9.0*
Kuder Occupational Preference Record	60	9.5*
Directions for Kuder	70	8.0
College G.E.D. No. 2	59	10.0*
Ohio State Psychological Test (Part 3)	37	15.0*
Strong Vocational Interest Blank	35	15.0*
Directions for Strong	73	7.5
Allport-Vernon	35	15.0*
Directions for Allport-Vernon	60	9.5*

* Starred grade scores represent Flesch's corrected grade placement for the area of extrapolation beyond the 7th grade.

Table 3

Grade and Reading Ease Level of Some Intelligence Tests as Determined by the Application of the Flesch Formula

	Directions		First Items		Middle Items		Last Items	
	R.E.	G.L.	R.E.	G.L.	R.E.	G.L.	R.E.	G.L.
1. A.G.C.T.	93	5.5						
a. Block counting			91	5.5	91	5.5	91	5.5
b. Arithmetic items			78	7.0	86	6.5	85	6.5
c. Vocabulary items			86	6.5	59	10.0*	62	9.0*
2. Henmon-Nelson, Form A, Grades 7-12	96	5.0						
a. Non-arithmetic items			86	6.5	56	11.0*	43	14.0*
b. Arithmetic items							96	5.0
3. Kuhlman-Anderson (Grade 6)	Oral							
a. Sample I			95	5.0			76	7.5
b. Sample II			95	5.0			76	7.5
4. Otis Higher Exam, Form C	81	7.0	80	7.0	71	8.0	50	13.0*
5. Terman-McNemar, Form D	86	6.5						
a. Information test			90	5.5	76	7.5	58	10.0*
b. Logical selection			64	9.0*	39	15.0*	10	17.0*
c. Analogies			87	6.0			80	7.0
d. Best answers			62	9.0*			63	9.0*

* Starred grade scores represent Flesch's estimated corrected grade placement for the area of extrapolation beyond the 7th grade. Reading grade placement scores indicated by Tables 2 and 3 derived from Flesch's (9, 11) conversion tables.

vocabulary ability indicates that knowledge of words is less important than reading ability as a whole.

A limited check on the accuracy of the Flesch formula was made by applying the Lorge (18) formula, to the Minnesota Multiphasic Personality Inventory and the Dale-Chall (8) formula to a sample of the information section of the Terman-McNemar Intelligence Test. The application of the Lorge formula to the same sampling of the MMPI resulted in a grade score of 6.1, which is in agreement with results on the Flesch formula. The Lorge formula is based on average sentence length, number of prepositional phrases per 100 words, and number of difficult words not appearing in Dale's list of 769 easy words. The application of the Dale-Chall formula to the Terman-McNemar Information Test gave a corrected grade level score of 12, which is somewhat higher than the grade score arrived

at for the same sample (last items) by the application of the Flesch formula. The Dale-Chall formula is based on sentence length and number of difficult words not appearing on Dale's list of 3000 familiar words.

Scrutiny of Tables 2 and 3 suggests tentative observations which may have implications of value to counselors if the results in the tables can be accepted as approximating the reading levels indicated. In some instances, test directions gave a reading grade score below the readability level of test items. Interest tests, particularly the Strong Interest Test and the Study of Values, appear to be at a reading level considerably higher than that attained by the mean of the high school or general population. The relatively high difficulty level of the Strong Interest Test, as reflected by the Flesch and Lewerenz formulas, seem to be in accord with the educational level of Strong's Occupational Criterion Groups, since

his groups are considerably above the mean of the general population. Strong (22) lists the average educational level of each occupational criterion group, and they range from 10.4 to 19.0. The mean educational attainment of 35 out of 39 of Strong's Occupational Criterion Groups was 14.5, and the standard deviation of the group was 2.4.

The Kuder Preference Inventory appears to have a reading level above that of the average junior high school student. Accordingly, it is of interest that a study by Christensen (7) indicated that 9th grade pupils of mean average intelligence had erroneous ideas concerning the meanings attached to 21 key words in the Kuder Preference Record. After instruction concerning the meaning of these terms, the results indicated that such instruction probably played a role in causing subjects to change their preferences.

The Bell and the MMPI personality tests have reading comprehension levels which would appear to be understood by most adults and junior high school groups.

Some of the commonly used intelligence tests measured appear to be at a reading level above that of the group for which the tests were designed and also appear to be spiral power tests which may to some degree be measuring reading skills. Center and Persons (6) found that after one semester of remedial reading instruction, some pupils gained from 12 to 17 IQ points on the Terman Group Test of Mental Ability. A class of 40 pupils made an average gain of 4.3 IQ points under the above conditions. The Otis and the Kuhlmann-Anderson seem to more nearly match the reading levels of the groups for which they were designed.

The block counting and arithmetic items of the AGCT appear to be at the reading level of adults whose mental age is 13 and above. These two subtests of the AGCT would be in accord with the reading level of most junior and senior high school students. The reading level of the entire test would not be too difficult for most senior high school students and above average adults, since the most difficult section has a reading level of about grade 10. However, it must be remembered that over one half of the adult population is below this level and may be mismeasured.

Some non-arithmetic items of the Henmon-

Nelson test and the logical selection of Terman-McNemar are at about the college reading level. That is, about grade 14.

Summary

Assuming that the above observations are generally correct, the findings of applying reading ease formulas to directions and items of commonly used tests may be summed up as follows:

1. Counselors, psychometrists, teachers and personnel technicians may have to re-evaluate their past and present practices regarding test selection, administration and interpretation.

2. Psychometrists and counselors cannot assume that if a client comprehends the directions for a test he will necessarily comprehend the test items.

3. There is need for interest inventories that junior high school students and the general population can easily comprehend.

4. The reading level of the Bell and MMPI tests appears to be well adapted to most of the general population and most junior high school groups. However, the variability in reading ability in these groups would indicate that even these tests would not adequately measure the clients in the lower end of the reading distribution.

5. The Army General Classification Test would appear to have possibilities for more extensive use at the junior and senior high school levels and for use with general population groups. It also has possibilities for use as a substitute for individual tests of mental ability. This conclusion is fortified by a recent study by Tamminen (23) which shows a correlation of .83 between the Wechsler-Bellevue and the AGCT.

6. There is need for group tests of mental ability that are not affected by an individual's abilities in reading. It appears that some of our commonly used intelligence tests tend to favor those individuals who have, because of their environment, attained a high degree of reading skill. If we are to encourage and assist in the development of an individual's natural potentialities, we need reliable means of measuring his real-innate intelligence, no matter how inadequate his reading skills may be. Our country's brain power is likely its

greatest resource and certainly every effort should be made to discover and develop it.

Received November 28, 1949.

References

1. Aukerman, R. C., Jr. Differences in the reading status of good and poor eleventh grade students. *J. educ. Res.*, 1948, 41, 498-515.
2. Baker, G., and Peatman, J. G. Tests used in Veterans Administration advisement units. *Amer. J. Psychol.*, 1947, 49, 99-102.
3. Berkshire, J. R., Bugental, J. F., Cassens, F. P. Test preferences in guidance centers. *Occupations*, 1948, 26, 337-343.
4. Betts, E. A. *Foundations of reading instructions*. New York: American Book Co., 1946. Pp. 445-485.
5. Bond, G. L. Identifying the reading attainments and needs of students. *Yearb. nat. Soc. Stud. Educ.*, 1948, 47 (2), 224-249.
6. Center, S. S., and Persons, G. L. *Teaching high school students to read*. New York: Appleton-Century Co., 1937.
7. Christensen, T. E. Some observations with respect to the Kuder Preference Record. *J. educ. Res.*, 1946, 40, 96-107.
8. Dale, E., and Chall, J. Formula for predicting readability instructions. *Educ. Res. Bull.*, Ohio State Univ., 1948, 27 (2).
9. Flesch, R. *Marks of readable style: a study in adult education*. New York: Bur. of Publ., Teachers Coll., Columbia Univ., 1943. (Contr. to Educ. No. 897.)
10. Flesch, R. *The art of plain talk*. New York: Harper and Brothers, 1946.
11. Flesch, R. A new readability yardstick. *J. appl. Psychol.*, 1948, 32, 221-233.
12. Flesch, R. *The art of readable writing*. New York: Harper and Brothers, 1949.
13. Gates, A. I. *The improvement of reading*. New York: The Macmillan Co., 1947. Pp. 631.
14. Gray, W. S. The nature and extent of the reading problems in American education. *Educ. Record Supplement*, 1938, 19 (11), 87-104.
15. Johnson, R. H. The problem of veterans' reading level in the counseling process. *Minnesota Counselor*, 1948, 3 (4).
16. Klare, G. R. Understandability and indefinite answers to public opinion questions. *Int. J. Opin. Attitude Res.*, 1950, 4, 91-96.
17. Klare, G. R. Evaluation of quantitative indices of comprehensibility in written communications. Unpublished Ph.D. thesis, 1950, Univ. of Minnesota.
18. Lorge, I. Predicting readability. *Teach. Coll. Rec.*, 1944, 45, 404-419.
19. Lorge, I., and Blau, R. Reading comprehension of adults. *Teach. Coll. Rec.*, 1941, 43, 189-198.
20. Monroe, M. *Children who cannot read*. Chicago: The University of Chicago Press, 1932. Pp. 10-12.
21. Steffire, B. The reading difficulty of interest inventories. *Occupations*, 1947, 26, 95-96.
22. Strong, E. K., Jr. *Vocational interests of men and women*. Stanford: Stanford Univ. Press, 1943. Pp. 694-702.
23. Tamminen, A. W. A comparison of the army general classification test and the Wechsler-Bellevue intelligence scales. Unpublished.
24. Witty, P. A. Current role and effectiveness of reading among youth. *Yearb. nat. Soc. Stud. Educ.*, 1948, 47 (2).

How Readable are Occupational Information Booklets?

Arthur H. Brayfield and Patricia Aepli Reed

University of California, Berkeley, California

Dullness and incomprehensibility have never to our knowledge been postulated as twin goals for writers of occupational information. Indeed, the injunction to make such information interesting and readily understandable has been sounded since the early days of the vocational guidance movement.

"It (occupational information) needs to be put in terms that can be used by all kinds and conditions of people . . . it should be made available in very simple and popular form" was the admonition of Charles R. Richards as early as 1915 (3, 514). In the mid-twenties Mary C. Schauffler wrote: "A study may be clear, definite, accurate, and brief and yet be dull reading, and on that account, may be of little value for classroom purposes. It is not enough that material be presented clearly and directly, it must also be presented in a manner interesting and understandable to the group that is to use it" (4, 136). More recently, in a comprehensive treatment of occupational information published in 1946, Shartle expressed the belief that "While information must be carefully checked for accuracy, it must also be arranged and written in a style that is readable and easy to use" (5, 76).

What are the facts? Have writers and publishers of occupational information succeeded in producing readable materials? The writers have discovered no studies bearing on the problem; this paper reports a preliminary investigation of the question "How readable is occupational information?"

Procedure

To answer this question we applied the revised Flesch method of measuring readability and human interest to sample passages from current occupational information literature.

The revised Flesch readability formulas were described at some length in 1948 (1). Formula A is essentially a test of level of abstraction and is thought to be an index of comprehension difficulty. Formula B predicts the effect of

two "human interest" elements on comprehension. Flesch considers its real value to lie in the fact that "human interest will also increase the reader's attention and his motivation for continued reading" (1, 226).

An attempt was made to sample widely the current occupational literature. Publications covering a variety of occupations and industries were included from many different publishing sources. A major difficulty encountered was the paucity of occupational information at the so-called lower job levels.

Reading ease and human interest scores were determined for these materials as follows: (a) Five samples were chosen from each piece of writing; (b) Each sample was marked off to include 100 words; and (c) The steps outlined by Flesch were taken to compute the readability and human interest scores.

Results

First, we analyzed 31 publications describing professional level occupations. The results are given in Table 1. The readability standards suggested by Flesch (1) were referred to for interpretation of the findings.

Flesch groups reading ease scores into seven categories ranging from "Very easy" to "Very difficult" with a further description according to representative magazine publications ranging from "Comics" to "Scientific." The most striking fact obtained from this analysis is that 65 per cent of the publications studied fall into the category "Very difficult" as represented by scientific journals and the remainder are classified as "Difficult" or equivalent to academic magazines.

With respect to human interest, Flesch describes five categories of scores ranging from "Dull" to "Dramatic." Typical magazines range from "Scientific" to "Fiction." Of these 31 publications 71 per cent are judged by the Flesch formula for human interest to be "Dull" and appropriate to scientific journals. The remaining 23 per cent are rated in the

Table 1

"Reading Ease" and "Human Interest" Scores for Publications, Arranged by Source,
Describing Selected Professional Level Occupations

Source	Occupation									
	Lawyer		Engineer		Pharmacist		Nurse		Teacher	
	"R.E."	"H.I."	"R.E."	"H.I."	"R.E."	"H.I."	"R.E."	"H.I."	"R.E."	"H.I."
U. S. Department of Labor	7	4	14	5	36	5	4	4	12	8
U. S. Employment Service	32	1	43	1	7	12	15	6	18	14
Science Research Associates	29	10	18	7	30	3	29	9	30	8
Institute for Research	31	13	11	7	37	2	26	10	21	8
Bellman Publishing Company	13	9	49	2	32	9	41	19	—	—
Occupational Index, Inc.	34	9	3	7	—	—	—	—	9	5
Professional Organizations	20	4	—	—	28	3	22	17	5	0

adjacent category "Mildly interesting" on a par with trade publications.

It might be hypothesized that these findings are accounted for by the complexity and technical nature of professional occupations. The findings of our study of a limited sampling of skilled and semi-skilled occupations lend only slight support to this contention.

The data of Table 2 indicate that these 19 publications are slightly more readable than those for professional occupations: only 53 per cent fall into the "Very difficult" or scientific classification for reading ease. However, none of the samples rated less than "Difficult" or academic, which is the next category on the Flesch scale.

Results for the human interest analysis follow the same pattern: 68 per cent may be

categorized as "Dull" or scientific. The remainder, with one exception, are comparable to trade journals or "Mildly interesting." The exception barely rates as "Interesting."

Industrial as well as occupational classifications were studied. The results for reading ease of these 16 publications dealing with entire industries are reported in Table 3. They are similar to the previous findings. Of these, 62 per cent rank as "Very difficult." An additional one fourth rated "Difficult," one was only "Fairly difficult" similar to quality magazines, none were "Standard," and in the remaining publication the hotel industry actually is described in a Science Research Associates pamphlet in a way which earns it a description as "Fairly easy," like slick-fiction.

Table 2

"Reading Ease" and "Human Interest" Scores for Publications, Arranged by Source,
Describing Selected Skilled and Semi-Skilled Level Occupations

Source	Occupation									
	Secretary		Plumber		Welder		Truck Driver		Domestic	
	"R.E."	"H.I."	"R.E."	"H.I."	"R.E."	"H.I."	"R.E."	"H.I."	"R.E."	"H.I."
U. S. Department of Labor	39	4	5	0	32	1	—	—	—	—
U. S. Employment Service	—	—	29	15	22	8	—	—	—	—
Science Research Associates	35	19	46	12	49	4	31	17	40	14
Institute for Research	31	20	20	9	—	—	—	—	—	—
Bellman Publishing Company	23	7	—	—	—	—	—	—	—	—
Occupational Index, Inc.	3	5	—	—	19	7	15	8	27	5
Commonwealth Book Company	—	—	1	2	31	4	—	—	—	—

Table 3

"Reading Ease" and "Human Interest" Scores for Publications, Arranged by Source, Describing Selected Industries

Source	Industry									
	Retail		Motion Picture		Real Estate		Hotels		Iron and Steel	
	"R.E."	"H.I."	"R.E."	"H.I."	"R.E."	"H.I."	"R.E."	"H.I."	"R.E."	"H.I."
Science Research Associates	52	13	23	16	25	10	76	16	41	3
Institute for Research	—	—	3	1	12	14	—	—	—	—
Bellman Publishing Company	8	14	31	15	—	—	23	14	30	7
Occupational Index, Inc.	16	4	—	—	10	9	26	20	—	—
Commonwealth Book Company	16	9	—	—	—	—	—	—	—	—
Western Personnel Institute	—	—	42	13	—	—	—	—	—	—

Human interest values are slightly more favorable for the potential consumer of these publications. Only 37 per cent rate at the extreme as "Dull." However, all but one of the rest fall into the adjacent "Mildly interesting" classification. The single exception is judged "Interesting" like digest magazines.

It was not a prime objective of this investigation to make comparisons among publishers with the one exception described in the next paragraph. It may be noted in passing though that there do not appear to be any significant trends. No differences of any magnitude in readability are suggested by these data although they might appear if the entire series of publications from each source were studied in a similar manner. Actually none of these publishers appears to be producing materials which begin to meet Flesch standards indicative of easy and interesting reading.

An analysis was made of the publications of business and industrial concerns themselves to determine whether or not the highly ballyhooed advertising practices of "big business" have influenced the readability of the occupational information supplied by such firms. The results of the study of the occupational information materials of twelve companies are reported in Table 4. These include such representative titles as "What Shell Means to You," "Opportunities for Employment," and "What about Your Future?"

Actually, 83 per cent rate as "Very difficult." The remaining two publications are split between "Difficult" and "Fairly difficult."

Two thirds are judged "Dull." One fourth are "Mildly interesting" and only one publication achieves a designation as "Interesting."

Summary

In all, 78 pieces of occupational information literature from 24 different sources were analyzed. Almost two thirds ranked as "Very difficult" or at the scientific level with respect to reading ease while another 32 per cent were ranked "Difficult."

Almost exactly the same proportions held for the categories "Dull" and "Mildly interesting" when human interest scores were studied.

Table 4

"Reading Ease" and "Human Interest" Scores for Occupational Information Publications of Private Business and Industry

Source	"Reading Ease"	"Human Interest"
Burroughs	20	5
Chase National Bank	29	15
Corning Glass Works	23	6
Equitable Life Insurance	14	9
General Motors	15	17
J. C. Penney	51	36
Proctor and Gamble	7	9
Roos Bros. Department Store	36	11
Shell Oil Company	23	9
Union Oil Company	19	9
United Air Lines	21	7
United States Steel	22	3

Fewer than five per cent of these publications reach the readability level of the popular "digest" magazines.

Within the limitations of the Flesch formulas for measuring readability and the limitations of our sampling of occupations, industries, and publishers we conclude that current occupational information falls far short of meeting the requirements for comprehension and interest which have been suggested through the years by persons intimately concerned with the preparation and use of such information. Dullness and incomprehensibility reign supreme. Writers and publishers of occupational information might well consult Flesch

(2) and others if they are seriously interested in "The Art of Readable Writing."

Received October 28, 1949.

References.

1. Flesch, R. A new readability yardstick. *J. appl. Psychol.*, 1948, 32, 221-233.
2. Flesch, R. *The art of readable writing*. New York: Harper and Brothers, 1949.
3. Richards, C. R. What we need to know about occupations. In M. Bloomfield (Ed.), *Readings in vocational guidance*. Boston: Ginn and Company, 1915. Pp. 504-514.
4. Schauffler, M. C. Standards for evaluating occupational studies for a critical bibliography. In F. J. Allen (Ed.), *Practice in vocational guidance*. New York: McGraw-Hill, 1927. Pp. 132-138.
5. Shartle, C. *Occupational information*. New York: Prentice-Hall, 1946.

Validity of an Emotional Key on a Short Industrial Personality Questionnaire *

Mildred B. Mitchell

VA Mental Hygiene Clinic, Ft. Snelling, St. Paul, Minnesota

and

Harold F. Rothe

Stevenson, Jordan & Harrison, Inc., Chicago, Illinois

This is the third in a series of papers dealing with the Objectivity score on a short personality questionnaire designed for industrial use. The questionnaire is restricted and is not available for general use, but the principles involved are more important than is the particular form. In the first paper (3) the questionnaire was described and the use of the Objectivity key, patterned after the L scale of the MMPI (2), was described. It was shown there that Emotionality scores, in particular, seemed to be related to Objectivity scores when the questionnaire was administered to industrial personnel.

The information presented in that paper and the use of that information as described is of no value unless it can be shown that the Objectivity key is a valid measure of "faking" and unless the other keys are also valid for measuring various personality characteristics. A second paper described an experiment with college students in which it was shown that the Objectivity key, and other keys could be "faked," if the respondents so desired (1).

Faking the questionnaire to "look good" resulted in low Emotional and also low Objectivity scores that gave the interviewer a clue that "faking" might be present. Thus it was concluded that the Objectivity key is a valid measure of what it is intended to measure. As described in the first paper, the Objectivity key is interpreted in industrial practice to "protect" highly objective persons rather than to locate "fakers."

The purpose of the present paper is to present evidence on the validity of the Emo-

tionality key. This key is shown to be a valid measure of maladjustment. Accordingly the validity of the procedure of using the Objectivity scores to determine the set of norms to use in interpreting the Emotionality key on a short personality questionnaire (50 items) in industrial screening is demonstrated.¹

The Criterion Group

For purposes of this study the personality questionnaire was administered routinely to a sample of 100 male patients consecutively admitted to the Mental Hygiene Clinic of a Veterans Administration Center. All of the respondents had been admitted to the clinic as patients, and this fact is used as the criterion of their maladjustment. The specific nature of their maladjustments is unimportant here because the Emotionality key under discussion is a screening key and does not yield responses that are diagnostically refined.

For comparative purposes a sample of 100 industrial employees was selected from master data sheets of test scores. Every fifth man was selected, beginning with the most recent entry on the data sheets and proceeding until a sample of 100 cases was obtained. The two groups, clinical and industrial, are accordingly not matched in any respect except sex.

Results

Complete distributions of the scores of these two groups together with the means and standard deviations on each of four keys (Objectivity, Emotionality, Social Dominance,

* Published with permission of the Chief Medical Director, Department of Medicine and Surgery, Veterans Administration, who assumes no responsibility for the opinions expressed or the conclusions drawn by the authors.

¹ The writers wish to thank Miss Judy Yackle for assisting in the preparation of this paper, and the Psychological Trainees, particularly Arthur Schomp and Harold Gilberstad, at the VA Mental Hygiene Clinic, for helping to collect the data on patients.

Table 1

Distributions of Objectivity Scores of 100 Clinical Patients and 100 Industrial Employees

Objectivity Score	Clinical Patients	Industrial Employees
0	5	5
1	19	20
2	22	23
3	26	22
4	18	18
5	6	10
6	4	2
Mean	2.7	2.7
S.D.	1.5	1.5

and Drive) are presented in Tables 1, 2, 3, and 4.

The Objectivity Key. The two groups were surprisingly well matched when their Objectivity scores were compared. The mean score for each group was 2.7 and the S.D. for each group was 1.5, as is shown in Table 1. The critical ratio of the differences between the means of these distributions is .05, which is a non-significant difference.

The Emotionality Key. The two groups were quite disparate on the Emotionality key, as shown in Table 2. The mean score for the

Table 2

Distributions of Emotionality Scores of 100 Clinical Patients and 100 Industrial Employees

Emotionality Score	Clinical Patients	Industrial Employees
0	0	0
1	1	7
2	3	11
3	2	22
4	5	11
5	9	18
6	8	8
7	11	8
8	18	6
9	14	9
10	14	0
11	7	0
12	6	0
13	2	0
Mean	7.9	4.6
S.D.	2.6	2.3

clinical sample was 7.9, with an S.D. of 2.6. The mean for the industrial sample was 4.6, with an S.D. of 2.3. The critical ratio between these two samples was 9.3, hence it is concluded that the difference is a real one. Since the clinical sample was maladjusted according to an outside criterion, and since the industrial sample was not known to be maladjusted, it is concluded that the Emotionality key is a valid measure of maladjustment.

The correlation between the Emotionality and the Objectivity scores for this clinical sample was .25, which is small enough to be of little or no practical significance. This is an important point since, as previously shown by a co-variation of distributions, these keys are rather highly related for industrial or college samples. For the industrial sample discussed in this paper, the correlation between these two keys was .52.

Thus it is shown that there is not an inevitable high degree of relationship between the Objectivity and the Emotionality scores obtained on this questionnaire. The fact that such a relationship does occur with industrial samples, and does occur when college students take the questionnaire and "fake" it, further strengthens the belief that low Objectivity scores are indicative of the existence of "faking." Stating this in terms whereby this form is used in actual practice, the belief is strengthened that persons who obtain high Objectivity scores and high Emotional scores in industrial interviewing situations should not necessarily be considered maladjusted. Rather, their scores should be interpreted on the basis of norms based upon persons of high Objectivity scores. Persons with low Objectivity scores and high Emotionality scores may be suspected of poor adjustment, requiring more intensive interviewing and testing procedures. The relationship of Objectivity scores to Emotionality scores does not appear to be a spurious one in industrial practice.

The Social Dominance and Drive Keys. On the Social Dominance scale the clinical patients obtained lower scores than did the industrial sample. These data are presented in Table 3. The clinical mean was 4.9, with an S.D. of 2.2. The industrial mean was 6.2, with an S.D. of 1.7. The critical ratio between the two

Table 3

Distributions of Social Dominance Scores of 100 Clinical Patients and 100 Industrial Employees

Social Dominance Score	Clinical Patients	Industrial Employees
0	0	0
1	8	2
2	11	1
3	9	2
4	16	12
5	13	16
6	16	17
7	16	27
8	6	18
9	5	5
Mean	4.9	6.2
S.D.	2.2	1.7

groups was 4.69, which indicates a significant difference.

On the Drive scale also the clinical patients were lower than the industrial sample as shown in Table 4. The clinical mean was 4.9 and the S.D. was 1.5. The industrial mean was 5.8 and the S.D. was 1.6. The critical ratio between the two groups was 3.52, which is again a significant difference.

Thus the clinical patients differed significantly from the industrial sample on three keys: Emotionality, Social Dominance, and Drive;

Table 4

Distributions of Drive Score of 100 Clinical Patients and 100 Industrial Employees

Drive Score	Clinical Patients	Industrial Employees
0	0	0
1	1	1
2	2	1
3	14	8
4	23	12
5	27	17
6	14	28
7	12	20
8	7	9
9	0	4
Mean	4.9	5.8
S.D.	1.5	1.6

but did not differ on Objectivity. The previous paper by Carr and Rothe (1), showed that students, when "faking to look good," obtained low Objectivity scores, low Emotionality scores, high Social Dominance scores, and unchanged Drive scores; when "faking to look bad" the students obtained high Objectivity scores, high Emotionality scores, low Social Dominance scores and low Drive scores. Thus the students, when "faking" could alter their scores on all three keys, but their extreme Objectivity scores revealed that they were "faking" and made possible an adjustment in the interpretation of their scores.

The clinical patients were apparently "normal" in their Objectivity, but low in Social Dominance and low in Drive. Accordingly, it is concluded that these patients as a group were actually low in those two characteristics. There is no reason to suspect that they were "faking" their scores on those two keys, particularly since they ranked on the low, and for many positions the undesirable, ends of the scales.

Conclusions

Data are presented in this paper showing that clinical patients who may be considered neurotic obtained significantly high scores on the Emotionality key of this short personality questionnaire, when contrasted with a random sample of industrial personnel. It is concluded from these data that this form is a valid screening device for indicating neuroticism. The form used here is not available for general use, since it has been developed by one company to aid its interviewing procedures. But the principles described here have broader implications. These are: (1) that a very short personality questionnaire for industrial purposes can be a valid screening device that will indicate neuroticism; and (2) that the use of an Objectivity key, similar to the L scale of the MMPI, can be a valid device that will aid in the interpretation of scores. It is apparent that other personality questionnaires can be made that will be as valid and as short as this one, and hence the restriction of the distribution of this form is actually not a serious handicap.

Received November 22, 1949.

References

1. Carr, E. R., and Rothe, H. F. Validity of an objectivity key on a short industrial personality questionnaire. *J. appl. Psychol.*, 1950, 34, 178-181.
2. Hathaway, S. R., and McKinley, J. C. *Manual for the Minnesota Multiphasic Personality Inventory*. New York: The Psychological Corporation, 1943.
3. Rothe, H. F. Use of an objectivity key on a short industrial personality questionnaire. *J. appl. Psychol.*, 1950, 34, 98-101.

Overall Job Success as a Basis for Employee Ratings

C. E. Jurgensen

Minneapolis Gas Company

The "halo" effect in merit rating is well known, and many different procedures have been devised to reduce its effect. In spite of such attempts, intercorrelations have remained so high as to cast doubt on the value of ratings on separate factors. In some cases intercorrelations have actually been higher than trait reliabilities (4). This has led some investigators (1, 5, 6) to recommend overall ratings variously called "performance on present job," "value to the company," "overall job success," etc.

Overall ratings lack the specific information found in trait ratings, but this objection is more theoretical than practical if intercorrelations between trait ratings are higher than their reliabilities. Overall ratings have certain advantages over trait ratings. They are apt to agree with the foreman's statements when he is questioned regarding an employee, and they are apt to agree with promotions, transfers, demotions, discharges and other personnel moves which are supposedly based on merit. Some persons condemn overall ratings on the basis that they are not analytical and may not be based on objective evidence. Yet these same persons may "validate" a trait scale on the basis of overall ratings or on promotions based on overall judgments. It would appear, then, that many persons believe that overall ratings are more valid than other means of determining employee merit and that some persons who disagree with this viewpoint in theory accept it in practice.

This report discusses two types of employee ratings based on overall job success.

Part I. Rank Order Merit Ratings

The rank order rating technique has received comparatively little attention in the literature of industrial psychology. The method consists of ranking employees on a specified continuum from best to worst. In the cases described here the continuum consisted of over-

all merit. The name of each employee was typed on a 3"X5" index card. The pack of name cards was given the rater with the information that names were in no particular order. Simultaneously, the pack of name cards was shuffled. Instructions were: "Please arrange these names in order from best to worst. Your best employee should be placed on top and the worst on the bottom. You can start from either end, or start from both ends and work toward the middle. You can make as many changes as you wish."

Rank orders were subsequently converted to scale scores advocated by Hull (3) and Guilford (2). Intercorrelation between the two types of scale scores was .998 with a total of 115 cases. Only the Hull method was subsequently used and reported here.

Case 1. Three foreladies ranked forty-four Inspector-Packers on an overall basis. Correlations between the raters were .88, .80, and .76. Stepped up reliability for the sum of the three ratings was .93.

Case 2. Three foreladies ranked twelve Inspector-Packers on an overall basis. Correlations between raters were .89, .77, and .72. Stepped up reliability was .92.

Case 3. Five foreladies ranked twenty-three Inspector-Packers on an overall basis. One month later the ranking was repeated without reference to the prior rankings. Correlations between the first and second rankings obtained for each forelady were .89, .88, .86, .76, and .73. Stepped up reliabilities for the sum of the two rankings by each forelady were .94, .94, .93, .86, and .85.

Intercorrelations were computed between the five raters. These are given in Table 1. Four of the five correlations lower than .80 involved forelady D who had been a forelady for less than one month and had had no previous rating experience or training. She was subsequently (and unrelated to these or other ratings) demoted to a non-supervisory position.

Table 1

Intercorrelations Between Five Foreladies Ranking
Twenty-Three Inspector-Packers

	A	B	C	D
A				
B	.75			
C	.84	.81		
D	.66	.65	.75	
E	.87	.82	.85	.70

Forelady B likewise had less than one month experience, but was subsequently considered a "good" forelady.

Case 4. Three inexperienced foreladies ranked thirty-four Inspector-Packers on an overall basis. Each forelady had been in supervisory work less than one month, and none had received any training in rating purposes or procedures. Intercorrelations between their ratings were .54, .53, and .19.

Case 5. An Assistant Sales Manager ranked fifty-two Salesmen on an overall basis of "value to the company as a salesman." To facilitate the ranking of a number as large as fifty-two, he was first asked to sort the name cards into three piles ("best," "average," and "worst") containing approximately the same number of names each. The three groups were then placed in rank order and merged. Changes of opinion in rank order were permitted until the judge was completely satisfied with his rank order. One month later the process was repeated. The Pearsonian correlation between the repeated ratings (scaled scores) was .94. Stepped up reliability for the sum of the two ratings was .97.

Rating reliability of this magnitude is so high as to be suspect. No evidence was found to indicate it was spuriously high or that similar results could not be obtained by other persons under the same conditions. It is believed that the high reliability obtained was due primarily to four factors: (1) Rank order merit ratings on an overall basis may inherently be more reliable than more frequently used procedures; (2) The rater was an exceedingly conscientious person and was highly motivated to be as accurate as possible in his ratings; (3) During the previous year the rater had re-

ceived almost twenty hours of individual training in rating purposes, principles, and procedures; and (4) The rater was well acquainted with his subordinates, having worked intimately with them for several years.

Conclusions for Part I

1. Ratings obtained from experienced supervisors were more reliable than those obtained from inexperienced supervisors.

2. Highest reliability was obtained from a supervisor who was highly motivated, had received individual training in rating, and was thoroughly familiar with the work of his subordinates.

3. Rank order ratings on an overall basis are simple to obtain and can have a high degree of reliability.

Part II. Multiple Item Scale for Rating Overall Job Success

The simple rank order method just described will not always be suitable. It requires that ratings be made for a group of employees. Sometimes a single employee must be rated, as when determining action to be taken on a probationary employee. Although rank order rating is not applicable, an overall rating may be desired. However, it is somewhat disturbing to base a decision on a single overall rating in view of the commonly accepted hypothesis that, other factors being equal, the reliability of a measurement is related to the number of items making up the measurement.

This section deals with an attempt to devise a four item scale in which each item consists of an overall rating of job success. Such a scale permits computation of split-half reliabilities for groups of employees and internal consistency for a single employee.

In developing a multiple item rating scale for overall job success it was considered important: (1) to word items in such way that their identicalness was not obvious; (2) to provide a varying number of rating categories so that raters would not automatically check the same category in each item; and (3) to provide two open end questions on each employee's strong points and weak points. A copy of such a scale is given in Figure 1.

Name of Employee _____ Position _____

Department _____ Rated by _____ Date _____

Place a check in the box which best indicates how satisfied you are with the present job performance of this employee.

☒ 68 Exceedingly well satisfied with employee	☒ 28 Definitely no
☒ 56 Well satisfied with employee	☒ 35 Probably no
☒ 45 Generally satisfied with employee	☒ 44 Probably yes
☒ 35 Somewhat disappointed with employee	☒ 56 Definitely yes
☒ 28 Quite disappointed with employee	
☒ 20 Very disappointed with employee	

If at the time this employee was hired you knew everything you now know about him, would you have recommended his employment?

What do you consider to be the greatest strong points of this employee?

Where would you grade this employee in a large group of persons holding this same job?

☒ 26 Lowest 10%	_____
☒ 34 Next 20%	_____
☒ 44 Middle 40%	_____
☒ 53 Next 20%	_____
☒ 63 Highest 10%	_____

Which of the following terms best describes the overall job performance of this employee?

☒ 70 Excellent, far exceeds job requirements	
☒ 60 Good, exceeds job requirements	
☒ 52 Average plus, slightly above job requirements	
☒ 43 Average, meets job requirements	
☒ 34 Average minus, slightly below job requirements	
☒ 28 Poor, partially meets job requirements	
☒ 17 Very poor, does not meet job requirements	

What do you consider to be the greatest weaknesses of this employee?

FIG. 1. Employee evaluation report including the composite normalized score weights based on total of 810 ratings.

Procedure. Twenty-one supervisors representing all divisions of the company filled in 405 Employee Evaluation Reports. The largest number of employees rated by a single supervisor was thirty-seven and the smallest number was four.

Ratings on each item were converted into normalized standard scores using the mid-point percentile method. These scores were ex-

pressed with a mean of 50 and a sigma of 10 and comprise the weights given each rating. One month after the first ratings were obtained, the same supervisors again rated the same employees. Standard score weights were computed as previously.

The normalized score weights for the first ratings of 405 employees were compared with those of the repeated ratings. The mean dif-

ference was less than .3. Figure 1 includes the composite normalized score weights based on the total N of 810. These weights apply strictly only to the raters and employees included in the study. The method of selecting raters and employees permit generalization throughout the company concerned, but weights cannot be assumed to be applicable to other companies.

Scale Reliability. Inasmuch as weights are expressed in terms of standard scores, weights of two or more items can be added directly. This permits computation of split half reliability where two items comprise one half and the remaining two items the other half. The reliability (stepped up, as usual, by the Spearman-Brown prophecy formula) was .94 for the first set of ratings (N=405) and likewise .94 for the second set of ratings (N=405).

Twelve supervisors each rated seventeen or more employees. Split half reliability was computed for each individual supervisor. These reliabilities ranged from .82 to .99 with a median of .96 for the first set of ratings. For the repeated set of ratings these reliabilities ranged from .87 to .98 with a median of .94.

Repeat reliability was .88 for the total of 405 employees. For individual supervisors who rated seventeen or more employees, the reliabilities ranged from .78 to .98 with a median of .92.

Four supervisors each rated one group of 36 employees, thus providing six intercorrelations. For the first set of ratings they ranged from .60 to .83 with a median of .71. For the repeated ratings they ranged from .67 to .84 with a median of .76.

The above reliabilities follow the usual pattern of highest reliability for split halves, next highest for repeated ratings, and lowest for ratings by different supervisors. All reliabilities appear high as compared with reliabilities typically obtained from ratings. This is particularly significant in view of the fact that no supervisor was given any training whatsoever in the use of this scale, and few, if any, had received training in the use of any rating scale. This procedure was deliberately followed to avoid possible difficulties resulting from training on an experimental scale which might subsequently be discarded. The assumption was made that if the scale was found

to be satisfactory without any training it would be even more so with training.

Item Reliability. Reliability of the scale as a whole is dependent on the reliability of each item in the scale. Item reliabilities were therefore computed for the 405 repeated ratings by means of contingency coefficients corrected for errors of grouping due to small number of classes. The coefficients reported here can be considered equivalent to the customary Pearsonian correlation. Repeat reliabilities of the four items (N=405) are: (a) .87; (b) .80; (c) .82; and (d) .90.

Item Consistency. Item reliability can also be interpreted on the basis of item consistency.

The percentage of employees (N=405) given identical ratings on two occasions on each of the items was: (a) 77.0%; (b) 64.9%; (c) 61.5%; and (d) 81.0%. Very few discrepancies greater than one step were found for any item, such discrepancies being: (a) 1.0%; (b) 2.0%; (c) 2.2%; and (d) .7%. Because of the varying number of steps in the scale items the above percentages are not strictly comparable. An index was therefore devised to express the number and extent of inconsistencies, in relation to the maximum spread.¹

The item indexes for repeated ratings (N=405) were: (a) .95; (b) .91; (c) .93; and (d) .93. Item indexes for ratings by different raters (N=432) based on all possible pairings of 36 employees rated by four supervisors were: (a) .94; (b) .88; (c) .92; and (d) .89. Item consistency appears satisfactory, and no one item stands out as appreciably superior or inferior to any other item.²

Item Intercorrelations. The four items com-

¹ The formula is:

$$C.I. = 1 - \frac{N_1 + 2N_2 + 3N_3 \dots + nN_n}{nN_{(0+1+2+3 \dots n)}}$$

where N = number of ratings with the step discrepancy indicated by the subscript, and n = maximum step discrepancy. The formula is at a maximum of 1.00 when ratings are completely consistent and at a minimum of .00 when each employee is rated as high as possible at one time and as low as possible at another time.

² Item b contained the phrase "next 20%" in two places. This proved deceptive to raters, and appears to have lowered the reliability and consistency of this item. It has subsequently been reworded to read "next lowest 20%" and "next highest 20%" and it is believed that the reliability of the item will be improved considerably.

prising the scale were intended to be four ratings of the same characteristic, namely an overall rating of job success. Split half reliability was found to be high, thus indicating that the two halves were equivalent. Further evidence on this point is obtained from intercorrelations of the four items. These range from .82 to .90 thus being essentially the same, and essentially the same magnitude as the repeat reliability of each item.

Conclusions for Part II

1. Without receiving any training in the use of the scale, twenty-one supervisors rated 405 employees on two occasions on a four item overall merit rating scale. Under these circumstances split half reliability was .94, repeat reliability was .88 and correlation of ratings by different supervisors ranged from .60 to .84 with *n*'s of 36 each.

2. Item reliabilities and consistency indexes were satisfactory for each of the four items.

3. Item intercorrelations were of essentially the same magnitude as repeat reliabilities.

4. It is reasonable to believe that the results would be even more favorable if supervisors were trained in the use of the rating scale. It is therefore concluded that the technique of multiple overall ratings is a promising and practical technique.

Received November 18, 1949.

References

1. Ewart, E., Seashore, J. E., and Tiffin, J. A factor analysis of an industrial merit rating scale. *J. appl. Psychol.*, 1941, 25, 481-486.
2. Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1936, pp. 250-251.
3. Hull, C. L. *Aptitude testing*. Yonkers: World Book Company, 1928, pp. 382 ff.
4. Jurgensen, C. E. Intercorrelations in merit rating traits. *J. appl. Psychol.*, 1950, 34, 240-243.
5. Lawshe, C. H., Kephart, N. C., and McCormick, E. J. The paired comparison technique for rating performance of industrial employees. *J. appl. Psychol.*, 1949, 33, 69-77.
6. Pond, M. Success of factory workers. *Occupations*, 1936, 14, 940-944.

A Comparison of the Terman-Miles M-F Test and the Mf Scale of the MMPI

Olga E. de Cillis and William D. Orbison

University of Connecticut

Vocational counseling presents many occasions to use both the Terman-Miles Attitude-Interest Analysis Test and the Minnesota Multiphasic Personality Inventory.¹ The use of these tests, however, has frequently revealed marked discrepancies in scores on the masculinity-femininity scales. These discrepancies, like the low correlations reported in the literature for other masculinity-femininity scales (10, 12), suggest the hypothesis that although the T-M and the MMPI tests may discriminate between the sexes, they are nevertheless measuring different aspects of masculinity-femininity. An opportunity² to test this hypothesis on a large group was provided when both tests were incorporated in a battery of tests being used provisionally in the School of Business Administration at The University of Connecticut.

Subjects

The subjects used in this study were 129 men and 50 women undergraduate students at The University of Connecticut. The 129 men were junior and senior students from the School of Business Administration enrolled in a course in industrial psychology. They ranged in age from 20 to 31; the median age was 24. The 50 women were freshmen and sophomores in the College of Arts and Sciences enrolled in an introductory course in psychology. Their ages ranged from 17 to 25; the median age of the women was 18.

Procedure

The tests used were the T-M, Form B and the group form of the MMPI. Test administration took place during class periods with a lapse of one to two weeks between each administration. Presentation of the tests was varied

in AB, BA order. The nature and purpose of the study were concealed in order to avoid the falsification to which masculinity-femininity tests are particularly susceptible (Kelly, Miles and Terman (11), Gough (6), Benton (1), and Meehl and Hathaway (15)).

The MMPI was machine scored; the T-M was hand scored. Since the chances for making errors in hand scoring the T-M are great, each test was hand scored by two different individuals. Errors revealed by differences in the two scorings were corrected.

Results

MMPI. The men in this sample obtain a score somewhat more feminine than that reported by Schmidt (17) for men in a non-college population ($M=18.0$) and somewhat more feminine also than the mean for the normed population. The manual reports a mean raw score of 20.5 (9, p. 11) with a standard deviation of 5.0; the present study has yielded a mean raw score of 24.6, standard deviation 4.86 (Table 1). The fact that the men in this study obtain a more feminine score is to be expected since it has been demonstrated that college men score above the published norms (Brown (2), Hathaway³ (2)).

The data for women, as far as mean raw score and dispersion are concerned, compare more favorably with the norms. The manual cites a mean raw score of 36.5 (9, p. 12) with a standard deviation of 5.0. In this study a mean raw score of 37.1 and a standard deviation of 4.30 were obtained. The data of this investigation accord with the norms as well as with the results of Verniaud (20) on a non-college population and with Brown (2), Hampton (8), Lough (12), and Nance (16) on college populations.

T-M. The mean raw scores and standard

³ Cited from personal communication in Brown (2, p. 542).

¹ Hereafter the Terman-Miles Attitude-Interest Analysis Test will be referred to as T-M; the Minnesota Multiphasic Personality Inventory as MMPI.

² Kindly provided by Dean Laurence J. Ackerman.

Table 1

Means and Standard Deviations of M-F Test Scores

Test		N	Mean Raw Scores	Standard Deviations
MMPI*	Men	129	24.6	4.86
	Women	50	37.1	4.30
T-M†	Men	129	73.7	43.96
	Women	50	-32.6	42.45

* For the MMPI larger raw scores mean greater femininity for both sexes. The manual provides a set of norms for each sex in which larger *T* scores indicate a deviation in the direction of the opposite sex pattern.

† For the T-M test, larger positive raw scores mean greater masculinity; larger negative raw scores mean greater femininity. In interpreting the percentile scores the larger the percentile for men the greater the masculinity; the larger the percentile score for women, the greater the femininity.

deviations for men and women on the T-M are presented in Table 1. Judged by norms presented in the manual (18, p. 8), the males are fairly typical. A mean of 67.4 and standard deviation of 47.65 reported for 130 college men on Form A compare well with the obtained mean of 73.7 and standard deviation of 43.96 of this study.

However, the women in this study prove considerably more masculine than are similar groups described in the manual (18, p. 5). The mean for college women on Form A is -60.8 with a standard deviation of 39.15 (18, p. 8). When both forms are averaged the median score is -65 (18, p. 13); the standard deviation which may be estimated from that table is 44.15. In this study a mean of -32.6 with a standard deviation of 42.45 was obtained (Table 1). Evaluated by the T-M manual,

the women of this investigation score at the level of M.D. or Ph.D. women.

Because the mean for women cited in the manual (18, p. 8) is not comparable with the mean of this study, a *t* test was computed. The obtained *t* value is 4.08 for 178 degrees of freedom. This difference is reliable beyond the .01 per cent level. Thus in regard to the T-M, the mean obtained for women in this study is significantly lower, hence more masculine than that of the normed population. The standard deviations, however, are about the same. The *F* ratio of the two variances is 1.17 which is not significant.

Differences between Men and Women on the T-M and MMPI. The differences and their significances for each test as a whole are shown in Table 2. All are significant beyond the .01 per cent level of confidence. No less significant are differences between the sexes on the separate exercises on the T-M.⁴ In each instance the result justifies the assertion that men and women score differently on these tests of masculinity-femininity—a finding that conforms with the literature (3, 4, 5, 19).

Correlations. Thus it has been established that both tests are measuring differences between the sexes (Table 2). Yet the findings in the literature concerned with comparisons of other masculinity-femininity scales as well as observations made by the authors prior to this investigation led to the hypothesis that the T-M and the MMPI, although measuring some aspects of masculinity-femininity, do not mea-

⁴ In this analysis as well as those that follow, exercises 2 and 7 of the T-M are omitted since there was almost no variability for either group. Attention is also called to the fact that these exercises are the least reliable of the T-M test (19, p. 60).

Table 2

Differences between Men and Women on the T-M and MMPI

Differences between Men and Women							
Group	Means						
	T-M	MMPI	T-M ₁	T-M ₃	T-M ₄	T-M ₅	T-M ₆
Men	73.7	24.64	- 4.25	+5.18	+30.34	+40.40	+2.11
Women	-32.6	37.10	-12.16	-4.54	+12.24	-19.00	-4.76
Difference	+106.33	-12.46	+ 8.21	+9.72	+18.10	+59.40	+6.87
D	7.15	.74	1.36	1.21	3.42	5.22	1.01
t	14.90	-16.82	6.04	8.03	5.29	11.38	6.80
P	All values are much less than .001						

Table 3

Correlations between Tests of M-F; Correlations between MMPI and Subtests of T-M

Group	N	rx_y	$rx_1y^\dagger$	rx_3y	rx_4y	rx_5y	rx_6y
Men	129	-.295***	-.034	-.100	-.214**	-.315***	+.078
Women	50	-.365**	+.134	-.305*	-.108	-.213	-.213
Difference	—	-.070	+.168	-.205	+.106	+.102	-.291

* Significant at the 5% level of confidence.

** Significant at the 1% level of confidence.

*** Significant at the .1% level of confidence.

† The subscripts refer to the sub-tests of the T-M. Subtests 1, 3, 4, 5, and 6 of the T-M were correlated with the total Mf scale of the MMPI.

The confidence limits for 129 cases are: 5% level = .172; 1% level = .204; .01% level = .285. For 129 cases the confidence limits are: 5% level = .279; 1% level = .361.

sure the same aspects to any great extent. This hypothesis has been verified by the low correlations of the present study (Table 3). For men the correlation between T-M and MMPI is $-.30$; for women it is $-.37$. Both correlations are negative, as should be expected in correlating raw scores, since a high T-M raw score means masculinity and a high MMPI raw score means femininity.

The significance of the departures of the regressions from linearity was tested by the technique of analysis of variance (14, 255-62) for two regressions: the regression of MMPI scores on the total T-M for the male group, and the regression of MMPI scores on the total T-M scores for the female group. In neither group does the appropriate F ratio approach significance. For these groups, it can be concluded, there is no significant departure from linearity. The other regressions were judged by inspection to be linear.

The correlations obtained in this study are comparable in magnitude with those cited in the literature for other masculinity-femininity scales. Nance (16) compared the masculinity-femininity scales of the Guilford-Martin, Strong and MMPI. The correlations obtained averaged $+.40$ for men and $+.21$ for women.

It may further be assumed that the several exercises comprising the T-M may themselves measure somewhat different aspects of masculinity-femininity. Therefore, an attempt should be made to determine what variables each of the separate exercises of the T-M has in common with the MMPI. This has been done by correlating exercises 1, 3, 4, 5, and 6 of the T-M with the MMPI (Table 3).

For women it is exercise three (information) which correlates most highly with the MMPI. For the male group exercises four (emotional and ethical response) and five (interests) yield the highest correlations. The correlations thus obtained have also been tested for significance by applying R. A. Fisher's z transformation (14, 123-24). The correlations which attain each level of confidence have been appropriately indicated in Table 3. Only the results of exercises 3, 4, and 5 will be of concern in the following analysis since the correlations for the other exercises do not differ significantly from zero.

The highest correlation for women, obtained between exercise three (information) and the MMPI, the authors are at a loss to explain, since the MMPI did not derive any items from this exercise of the T-M (Table 4). This correlation, however, is significant at only the 5% level of confidence and the difference from zero may have occurred by chance. For men, the two exercises on the T-M that correlate most highly with the MMPI are exercises four and five. As far as exercise five is concerned, this finding is understandable since the bulk of the items that were used in the MMPI were derived from this exercise (Table 4).

Hathaway and McKinley report (9, p. 5), in discussing the choice of items for the Mf scale, that "some were inspired by Terman and Miles, and others are original." A total of 43 items on the T-M were judged to be identical or similar to 31 items on the MMPI (Table 4). The following items may serve to illustrate this agreement: (1) T-M, "Are your

Table 4

Analysis by Subtests of Items on the T-M
that Occur on the MMPI

Exercise on T-M*	No. of Identical or Similar Items on MMPI**
4. Emotional and Ethical Response	8
5. Interests	25
6. Personalities and Opinions	2
7. Introverted Response	8

* Exercises 1-3 inclusive on the T-M have been omitted from this table since they did not contribute any items to the MMPI.

** In this category are included either identical items that occurred in Form A of the T-M or items that were similar in content but not in format that occurred in Form B.

feelings often badly hurt?"; MMPI, "My feelings are not easily hurt"; (2) T-M, "There is plenty of proof that life continues after death"; MMPI, "I believe in a life hereafter"; (3) T-M, "Were you ever fond of playing with snakes?"; MMPI, "I do not have a great fear of snakes."

Of the 43 items on the MMPI derived from the T-M, 25 are from exercise five (interests). On the basis of "common elements," therefore, it is to be expected that the correlations with exercise five would be highest. Referring to Table 3 it will be noted that this correlation is highest for the male group but not for the female group. However, the magnitudes of the correlations for either sex are about the same, both are in the same direction, and they do not differ significantly when tested.

Discussion

It is evident that the T-M and the Mf scale of the MMPI do not show a high correlation. The correlations are significantly different from zero, however, a result which does indicate that the two tests are measuring something in common. On the other hand the correlations are low enough to suggest that the overlap is not great—a finding substantiated in the correlational study made by Nance (16) with other tests of masculinity-femininity. Much of what the two tests have in common may be ascribed to exercises four and five.

Although the correlations between the two tests are not high and do not appear to be

measuring the same thing, the highly significant differences between men and women obtained would indicate that both tests are validly measuring some *differentiae* of the sexes. These findings are understandable if masculinity-femininity comprises not one but many dimensions. Accordingly, the tests used in this investigation would be measuring somewhat different dimensions. The possibility of several masculinity-femininity factors has already been indicated, although not conclusively, by Guilford and Guilford (7), and Martin (13). A factor analysis of the T-M test determining the common factors might prove more fruitful. Until some such determination of what these scales are measuring, caution should be used in interpreting these results in counseling.

Summary

Fifty female and 129 male undergraduate students at the University of Connecticut were given both the Terman-Miles M-F Test and the Mf scale of the MMPI. Both tests clearly differentiated between the sexes but did not correlate very highly with each other. It is proposed that both tests are validly measuring different aspects of masculinity-femininity interests and attitudes. A factorial analysis of the Terman-Miles M-F test is suggested. Caution in using the scales in counseling situations is recommended.

Received August 18, 1950.

Early publication.

References

1. Benton, A. L. The Minnesota Multiphasic Personality Inventory in clinical practice. *J. nerv. ment. Dis.*, 1945, 102, 416-420.
2. Brown, H. S. Similarities and differences in college populations on the multiphasic. *J. appl. Psychol.*, 1948, 32, 541-549.
3. Burger, F. E., Nemzek, C. L., and Vaughn, C. L. The relationship of certain factors to scores on the Terman-Miles Attitude-Interest Analysis Test. *J. soc. Psychol.*, 1942, 16, 39-50.
4. Disher, D. R. Regional differences in masculinity-femininity responses. *J. soc. Psychol.*, 1942, 15, 53-61.
5. Gilkinson, H. Masculine temperament and secondary sex characteristics: a study of the relationship between psychological and physical measures of masculinity. *Genet. Psychol. Monogr.*, 1937, 19, 105-154.

6. Gough, H. D. Diagnostic patterns on the Minnesota Multiphasic Personality Inventory. *J. clin. Psychol.*, 1946, 2, 23-37.
7. Guilford, J. D., and Guilford, R. B. Personality factors S, E and M, and their measurement. *J. Psychol.*, 1936, 2, 107-127.
8. Hampton, P. J. The Minnesota Multiphasic Inventory as a psychometric tool for diagnosing personality disorders among college students. *J. soc. Psychol.*, 1947, 26, 99-108.
9. Hathaway, S. R., and McKinley, J. C. *Manual for the Minnesota Multiphasic Personality Inventory*. Minneapolis: Univ. of Minnesota Press, 1943.
10. Heston, J. C. A comparison of four masculinity-femininity scales. *Educ. and Psychol. Meas.*, 1948, 8, 375-387.
11. Kelly, E. L., Miles, C. C., and Terman, L. M. Ability to influence one's score on a typical pencil and paper test of personality. *Character and Pers.*, 1936, 4, 206-215.
12. Lough, O. M. Teachers college students and the Minnesota Multiphasic Personality Inventory. *J. appl. Psychol.*, 1946, 30, pp. 241-247.
13. Martin, H. G. The construction of the Guilford-Martin Inventory of Factors GAMIN. *J. Psychol.*, 1936, 2, 107-127.
14. McNemar, Q. *Psychological statistics*. New York: John Wiley and Sons, Inc., 1949, pp. 364.
15. Meehl, P. E., and Hathaway, S. R. The K factor as a suppressor variable in the Minnesota Multiphasic Personality Inventory. *J. appl. Psychol.*, 1946, 30, 525-564.
16. Nance, R. D. Masculinity-femininity in prospective teachers. *J. educ. Res.*, 1949, 42, 658-666.
17. Schmidt, H. O. Test profiles as a diagnostic aid: The Minnesota Multiphasic Personality Inventory. *J. appl. Psychol.*, 1945, 29, 115-131.
18. Terman, L. M., and Miles, C. C. *Manual of information and directions for use of Attitude-Interest Analysis Test*. New York: McGraw-Hill Book Company, Inc., 1938.
19. Terman, L. M., and Miles, C. C. *Sex and personality*. New York: McGraw-Hill Co., 1936, pp. 600.
20. Verniaud, W. M. Occupational differences in the Minnesota Multiphasic Personality Inventory. *J. appl. Psychol.*, 1946, 30, 604-613.

A Follow-Up Study on Satisfaction with Nursing

Helen Nahm

Duke University

During the Spring of 1947 a study¹ of the satisfaction of three groups of students was made at the Duke University School of Nursing. The groups consisted of 70 seniors (class of 1947), 62 juniors (class of 1948), and 52 freshmen (class of 1949). To measure satisfaction an adaptation of the Hoppock Job Satisfaction Scale² was used. To measure factors associated with satisfaction students were asked to respond to a number of questionnaire items designed to obtain information about their reactions to the situation in the school of nursing.

Findings of the study indicated that the freshmen students, at the end of a period of nine months in the school, were an enthusiastic and highly motivated group. Junior and senior students, on the other hand, were much less satisfied and showed many evidences of tension and frustration. Responses of the latter groups to questionnaire items indicated a genuine concern about the lack of time to give satisfactory care to patients; as well as to study, sleep, and participate in social and recreational activities. When asked to give suggestions for improvement these students recommended shorter and better-planned hours of work; the employment of a larger number of staff nurses so that patients might be given better care; more competent head nurses and supervisors; and a counselor to assist students with their personal, social, and emotional problems. They felt that courses should be made more interesting, and that methods of evaluating progress of students in their learning experiences on hospital divisions should be improved.

In the year following the publication of findings of this study a number of changes were made at the Duke University School of Nursing. Hours per week of class work and clinical experience on hospital divisions were reduced from 48 to 44. A full-time counselor and a

number of more highly qualified individuals were appointed to the faculty. Head nurses and supervisors were encouraged to take courses in ward management and teaching, and personnel work in schools of nursing. A more diversified program of social and recreational activities was provided. Every possible effort was made to create an environment in the nurses' residence conducive to the welfare and happiness of students. Faculty members worked with individual student leaders and with representative student groups to gain a better understanding of student needs and problems, as well as to give active assistance in planning for and making changes which seemed indicated.

Follow-up Study

To determine whether there is a relationship between environmental changes in a school of nursing and attitudes of students toward their experiences in the school, follow-up studies were made of the freshman students who participated in the original study (class of 1949) and also of the group of students who entered the school of nursing during the fall of 1947 (class of 1950). The Nursing Satisfaction Blank was administered to the two groups during the Spring of 1948, and again during the Spring of 1949.

Changes in satisfaction of the class of 1949 from the Spring of 1947 to the Spring of 1949 are given in Table 1, and changes of the class of 1950 from the Spring of 1948 to the Spring of 1949 in Table 2. Changes in mean scores and standard deviations for the two groups are given in Table 3.

For the class of 1949 the correlations between satisfaction scores for successive years are as follows: (a) Between 1947 and 1948 scores = 0.54; (b) Between 1948 and 1949 scores = 0.53; and (c) Between 1947 and 1949 scores = 0.41.

The differences between means of 1947 and 1948 scores ($t = 5.59$) and 1947 and 1949 scores

¹ Nahm, Helen. Satisfaction with nursing. *J. appl. Psychol.*, 1948, 32, 335-343.

² Hoppock, R. *Job satisfaction*. New York: Harpers, 1935.

Table 1

Changes in Satisfaction with Nursing from 1947 to 1949 of Students Enrolled in the Class of 1949

	1949					Per Cent
	Doesn't like it	Indifferent	Likes it	Enthusiastic	Total	
1947						
Enthusiastic		5	10	8	23	51.1
Likes it		4	13	3	20	44.5
Indifferent			2		2	4.4
Doesn't like it					0	0.0
Total	0	9	25	11	45	
Per cent	0.0	20.0	55.6	24.4		100.0

Table 2

Changes in Satisfaction with Nursing from 1948 to 1949 of Students Enrolled in the Class of 1950

	1949					Per Cent
	Doesn't like it	Indifferent	Likes it	Enthusiastic	Total	
1948						
Enthusiastic			8	11	19	47.5
Likes it			13	5	18	45.0
Indifferent		1	2		3	7.5
Doesn't like it					0	0.0
Total	0	1	23	16	40	
Per cent	0.0	2.5	60.0	37.5		100.0

Table 3

Changes in Mean Scores and Standard Deviations on the Nursing Satisfaction Scale of Two Groups of Students of the Duke University School of Nursing

	Class of 1949 (N = 45)		Class of 1950 (N = 40)	
	Mean	S.D. of Dist.	Mean	S.D. of Dist.
Spring 1947	23.2	1.96		
Spring 1948	21.4	2.44	23.2	2.18
Spring 1949	21.8	2.44	23.0	1.82

($t=3.80$) are significant at the 1 per cent level (in the direction of less satisfaction). The difference between means of 1948 and 1949 scores is not statistically significant ($t=1.30$).

For the class of 1950 the correlation between 1948 and 1949 mean scores is 0.54. The difference between means is not statistically significant. From these findings it seems evident that, for the class of 1949, there was a significant decrease in satisfaction from the first to the second year, and no significant increase from the second to the third year. For the class of 1950 the slight decrease in satisfaction

from the first to the second year is not statistically significant.

Factors Associated With Satisfaction

Differences in percentages of students of the class of 1949 responding to various questionnaire items designed to discover factors associated with satisfaction in nursing indicated the changes which had taken place since the time of admission to the school. Percentage differences (1947 and 1949) which are significant at the 1 per cent level indicate that students:

- were more likely to enjoy working with doctors and to feel that doctors approved of their work;
- more often had opportunities to express their ideas on hospital divisions;
- were less likely to enjoy living in the nurses' residence; and
- were less likely to enjoy formal classes.

Percentage differences which are significant at the 5 per cent level indicate that students:

- more often had opportunities to use their initiative on hospital divisions;
- were less likely to feel that the teaching program on hospital divisions was adequate; and
- were less likely to feel they lacked social skills needed to feel at ease in social situations.

Attitude Changes

For the class of 1949 (admitted in September 1946) typical statements which indicate attitude changes which took place after admission to the school are given as follows:

- I like nursing better, now that I have had more experience.
- I have a better understanding of the profession and its responsibilities.
- I feel more self-assurance when I care for sick people.
- I understand patients and others better.
- The thrill has worn off.
- I realize the many things a nurse needs to know in order to be proficient.
- I am more realistic and less sentimental about nursing.
- I have become more hardened to illness.

For the class of 1950 (admitted in September 1947) typical statements which indicate attitude changes are given as follows:

- I am more interested and enthusiastic than during the first year.
- I am less idealistic than when I entered.
- I have developed greater understanding of people.
- I derive deep satisfaction from the work.
- I didn't realize it involved so much work and study.
- I have matured during the past year.
- I have gained more self-assurance.
- I am better able to accept responsibility.
- I like it, but am not as enthusiastic as when I entered.

Eighty-five per cent of the class of 1949 and 87 per cent of the class of 1950 said that, if it were possible to go back a few years, they would again enter the Duke University School of Nursing. The remainder of the students either were undecided, or said they would probably enter a school nearer home.

Future Plans of Students

Eighty-one per cent of the class of 1949 said they would prefer institutional nursing after graduation. Fifty-two per cent said they would like to do general staff nursing and the remainder either teaching, administration, or supervision. Sixty per cent felt they would need additional preparation in some field of nursing after graduation. Forty per cent believed that experience as a general staff nurse was all that was needed.

When asked what they would like to be doing ten years from now, 90 per cent of the class of 1949 said they would like to be married. Only 13 per cent of the group would plan to continue in some field of nursing after marriage.

Summary

The fact that, for the class of 1949, there was a significant decrease in satisfaction from the first to the second year without a corresponding increase from the second to the third year would seem to indicate that, once students have lost their initial enthusiasm for nursing, it is not easily regained. On the other hand,

the fact that, for the class of 1950, there was no significant decrease in satisfaction with nursing from the first to the second year would indicate that it is possible for students to retain their initial enthusiasm for nursing.

Responses of students of the class of 1950 as to their attitude changes following admission to the school seem, in general, more favorable than those of the class of 1949. They indicate that some of the concomitants of a satisfactory nursing school program are greater understanding of self and others, deep personal satisfaction from the work itself, and a maturing of the entire personality structure. Responses of both groups indicate that students become less idealistic about nursing and more realistic as they progress through the school. This is probably both inevitable and desirable.

The statements which students make as to their attitude changes indicate that only a few become actively aware of the great public need for the future contribution which they may make. The fact that only a small minority hope to continue in nursing after marriage would tend to support this conclusion.

From both initial and follow-up studies of satisfaction of students of the Duke University School of Nursing the general conclusion may be drawn that there is an association between changes in the total environmental situation and the extent of satisfaction with nursing. It seems likely, however, that a period of from two to three years is required to change attitudes which are primarily negative to those which are predominantly positive. Students who have been very dissatisfied continue to be suspicious of the motives of individuals responsible for a school, even though these individuals make changes which are much desired. When the more advanced groups of students in a school are dissatisfied and unhappy, the attitudes of less advanced groups are inevitably affected after a period of time. However, as the satisfaction of each successive group of students increases, the morale of the entire student group undergoes a gradual change. Suspicion gives way to confidence and trust, and an atmosphere in which each student may grow and develop is finally created.

Received December 12, 1949.

Attitudes of Veterans toward Vocational Guidance Services

Frederick J. Gaudet, A. Ralph Carli and Leland S. Dennegar

Stevens Institute of Technology

The problem of the evaluation of guidance services is in the forefront of educational and psychological thinking today. Certainly, at no time in the past has so much money been spent on educational and vocational guidance nor at any time have so many highly trained psychologists and guidance counselors been employed in this field.

The attempts to get a picture of the efficacy of this vast program have been very limited. There has been only one published study which used a control group.¹ The Central Office of the Veterans Administration (VA), however, has in its files the results of many attitudinal surveys of veterans who have received vocational guidance in centers under its auspices. None of them, however, makes use of an adequate sampling.²

A similar study was conducted during this same period by the Service Center of the Psychological Corporation.³ This study dealt with a non-veteran population and many of its questions have been incorporated in the questionnaire on which the present study is based. These items were included for the purpose of comparing veteran with non-veteran guidance services in an effort to answer a question posed by many psychologists as to the efficacy of the VA guidance program. Doubts have been raised as to the value of a program which was so huge that its inception was almost an emergency and for which the limitation of funds per man implied a certain restriction on the qualifications of counselors who could be employed.

The Psychological Corporation had counseled 1,184 employees in a large industrial concern, many of whom would be transferred

or released at the termination of hostilities. An anonymous questionnaire was circulated to all of these employees asking for their evaluation of, or satisfaction with, the guidance process. Replies were received from 685 men and women.

The present study was based on the results of a questionnaire sent to 200 veterans who had received counseling at the VA Guidance Center at Stevens Institute of Technology. The survey was designed to get a picture of the attitudes of veterans toward the guidance program as a whole, and also to get opinions on specific phases of the procedure with a view toward improving the services.

The subjects were selected at random from the files of those who had received guidance between September 29, 1945 and February 16, 1946. All of these veterans had visited the center at least one year prior to the time the survey was started. The questionnaire, along with a letter explaining the purpose of the study, was mailed to these 200 men, of whom 81 answered. Those who did not answer the first letter were sent a second which produced an additional 51 responses, and a member of the staff interviewed those who did not answer either letter. Altogether 164 replies were secured. The remaining 36 did not answer the questionnaire for the following reasons: (1) Address changed, no forwarding address, 23; (2) Living outside of New Jersey, 5; (3) Re-enlisted, 2; (4) In VA hospital, 2; (5) Non-existent address, 1; (6) Died, 1; (7) Could not recall his impressions, 1; and (8) Refused to answer, 1.

With the exception of the man who refused to answer, and perhaps the one whose recollection was poor, it is apparent that probably no selective factor operated to distort the answers.

The following questionnaire was used:

Authorization No.

(Draw a circle around the word which is your answer to each of the questions on this page)

¹ Dech, A. O., and Reeves, P. Effects of advisement upon continuation in training under P.L. 346. *Sch. & Soc.*, 1948, 67, 429-431.

² Gaudet, F. J. The Veterans Administration Advisement and Guidance Program. *Sch. & Soc.*, 1949, 69, 251-254.

³ Anderson, Rose G. Reported and demonstrated values of vocational counseling. *J. appl. Psychol.*, 1949, 33, 460-473.

- 1.* As a result of your visit to this center did you get a better idea
 - a. of your strongest abilities?.....yes no doubtful
 - b. of your less strong abilities?.....yes no doubtful
- 2.* Did your results show you had underestimated your aptitudes and abilities
 - a. for particular jobs?..yes no doubtful
 - b. in general?.....yes no doubtful
- 3.* Did the guidance and counseling you received here
 - a. increase your self-confidence?.....yes no doubtful
 - b. decrease your self-confidence?.....yes no doubtful
- 4.* Did the guidance and counseling on the whole give you a better understanding of yourself?.....yes no doubtful
- 5.* Do you feel that your guidance and counseling was a worthwhile experience?
 - yes no doubtful
- 6.* Would you recommend a similar counseling center for non-veterans at their own expense?.....yes no doubtful
7. On which floor were you treated best?
 - 1st 2nd 3rd
8. On which floor were you treated least satisfactorily?.....1st 2nd 3rd
9. What part of our job do you think we do best?
 - 1st 2nd 3rd
10. What part of our job do you think we do poorest?
 - 1st 2nd 3rd
11. We improve our work through criticism and suggestions. What can you suggest that we can do to change or improve our services?
 - 1st 2nd 3rd
- 12.* At what age do you think guidance and counseling should be provided?
 - 1st 2nd 3rd

The first five items were designed to determine the effect of counseling on the veteran's view of himself. Item 6 was directed toward obtaining a general reaction to the counseling service. As a basis for self-criticism and an aid in improving Stevens service, items 7 through 10 provided an opportunity for evaluation of specific aspects of the program; concrete suggestions for improvement were requested in item 11. Item 12 solicited the veterans' opinions regarding the persistent

* Items marked with an asterisk were taken from the Psychological Corporation questionnaire. Some were modified slightly.

question "When should one receive guidance?" Of course, the correct answer is "continuously," but until our schools and colleges are able to furnish educational and vocational guidance, this is merely a theoretical answer.

The median age of the group was 25 years. Unfortunately, data on educational levels were not available for these men but a study of a similar sampling of 200 men who received counseling during the same period indicates that the median of the highest grade completed was 11 with lower and upper quartiles of 10 and 12, respectively.

Although responses were received from 164 veterans, not all answered every question. This failure to answer was particularly frequent in *b* of items 1, 2, and 3, and in items 8 and 10, in which the veteran thought he had answered the query in previous questions, or parts of questions, or presumably had nothing derogatory to say.

The answers to the first six items of the questionnaire are indicated in Table 1. In general, the results of the Psychological Corporation and Stevens studies are fairly comparable. It should be remembered in interpreting these data that the Psychological Corporation survey was based on a 58 per cent return and that no study was made of those who did not answer. The Stevens data were based on a return of 82 per cent, and a follow-up which indicated that those who did not respond were probably not a selected group.

The results for part *a* of item 1 indicate that the two centers were similar in acquainting the individual with his assets, but part *b* shows that the Stevens Center did not stress the man's handicaps or liabilities as much as the Psychological Corporation. Whether this emphasis on "less strong abilities" is desirable or not is a question of the philosophy of guidance, namely, whether guidance should give the client a picture of both his assets and his liabilities, or whether it should stress assets. Of course, both should be evaluated by the counselor, and the client should be told of his liabilities if he mentions educational or vocational objectives in which he would be handicapped.

The responses to item 2*b* indicate that less than one-third (30%) had under-estimated their abilities *in general*. However, it is inter-

esting to observe that considerably more (44%) had under-estimated their aptitudes and abilities for particular jobs. Again, the responses in the studies of the Psychological Corporation and Stevens seem to agree.

Probably as important as obtaining a clear picture of his aptitudes and abilities, is the effect of guidance on the individual's confidence in himself. The answers to item 3a indicate that 73 per cent of the veterans increased their self-confidence as a result of going through guidance, while only four per cent stated that their general self-confidence was lowered. It is probable that a good guidance procedure would increase the self-confidence of most individuals but decrease that of a few whose qualifications for particular jobs are questionable. Whether the "right" ones had their self-confidence increased or decreased, these data do not tell us. The differences between the two groups, veteran and non-veteran, are not significant.

The answers to item 4 were intended to give an over-all picture of what the guidance process does to an individual in helping him see himself realistically. The results of the questionnaires of the two centers are again similar; about three-quarters felt they had a better understanding of themselves.

Item 5 is probably the key question in the whole questionnaire. The objective of an educational and vocational guidance center is not primarily to give the individual "a better understanding" of himself (item 4), but to give him an understanding of himself in relation to his educational and vocational future. This dual function is probably reflected in the considerably higher percentage of (95 per cent) "yes" answers to item 5 than to those (74 per cent) to item 4. The percentages of "yes" answers to the two questionnaires, Stevens and Psychological Corporation, are almost identical.

Item 6 constitutes another method, a less direct one, of getting an over-all evaluation of what these veterans thought of the guidance program. It is possible that someone who had been through guidance would consider it desirable for himself but would not think it desirable or necessary for others. In both the Psychological Corporation and Stevens studies, the services were free. The question

asked was whether they would recommend it for others who would have to pay. It will be observed that in both studies there is a decrease in the percentage of "yes" answers as compared to item 5, but the difference is not large.

In comparing the two studies it should be noted that the Psychological Corporation study was based on a population who went through the guidance process voluntarily. The Stevens study included some veterans (22) who came to the guidance center because they asked for this service under Public Law (PL) 346. The majority, however, were PL 16 cases, disabled veterans (142) who were *compelled* to go to a VA center before they were permitted to take training under the VA. Of course, it is not implied that all disabled veterans came under compulsion. In fact, their disabilities may have caused them to want guidance more than the average individual.

Items 7 and 8 were used in the questionnaire in an attempt to "fractionate" the favorable or unfavorable attitudes shown by the veterans. The first floor of the Stevens Guidance Center building was staffed by VA personnel; counseling took place on the second and third floors, and testing was done on the third floor. One interesting feature of the replies is that although the questionnaire did not offer an opportunity to answer "all" to either of these questions, 25 per cent wrote in this answer to item 7, and only two per cent to item 8. Discontent was more frequently located on the

Table 1

Reaction of Veterans to Vocational Guidance

Item	Per Cent Answering:			
	Yes*	No	Doubtful	No Reply
1a	76 (76)	13	9	2
1b	43 (69)	9	15	33
2a	44 (51)	34	13	9
2b	30 (27)	35	11	23
3a	73 (80)	12	13	2
3b	4 (6)	60	9	27
4	74 (80)	13	13	1
5	91 (90)	4	5	1
6	84 (71)	7	8	1

* Percentages in parentheses are those obtained by the Psychological Corporation in its questionnaire.

first floor as indicated* in the low (5 per cent) percentage of "yes" answers to this floor in response to item 7 and the high (28 per cent) percentage of "yes" to item 8. To determine whether the least favorable attitude toward the first floor in response to items 7 and 8 had influenced the feeling toward the whole guidance procedure, answers to items 7 and 8 were correlated with items 4 and 5. The results indicated that there was no inter-relationship.⁴

Items 9 and 10 were included in the questionnaire to evaluate various parts of the guidance process. It is interesting that the most frequent answer to both questions was "all good"—38 per cent to item 9 and 25 per cent to item 10. The next most frequent favorable answer to item 9 (29 per cent) and the most frequent unfavorable answer to item 10 (14 per cent) was *counseling*. Since counseling might be considered the climax, or the part of the guidance program requiring the greatest professional skill, the frequency of these answers is noteworthy in that they were able to recognize its importance to them. The other answers to item 9 in order of frequency were: testing, "no reply," interviewing, amount of time devoted to the guidance process, and VA representatives. The other dissatisfactions expressed in answers to item 10 in order of frequency were: "no reply," amount of time devoted to the counseling, testing, VA representatives, lack of placement aids, "nothing good," and location of the center.

Cues as to the above favorable and unfavorable evaluations are indicated by the responses to item 11. A tabulation of these responses is presented in Table 2. Certainly, these responses give the impression that the respondents had taken the questionnaire seriously. No comments appeared to be facetious. The chief value of the comments has been to make the counselors more keenly aware of their need to learn what the counselee seeks from guidance. Analysis of the com-

⁴ A later study of a new group (200) of veterans who went through the guidance center when it was staffed with different VA personnel indicated that the answers to items 7 and 8 were not due to VA procedures but to the personnel carrying out these procedures. The answers to these items for "first floor" were 23 per cent and six per cent, respectively, in the second study.

Table 2
Veterans' Suggestions for Improving the Guidance Process

1. Increase length of process: more counseling (11), more tests (7), more school information (4), more job information (3), more information on veterans' benefits (1)	26
2. Decrease length of process: shorter program (17), fewer tests (7)	24
3. Do placement work: find employment (2), follow up to see that veterans get jobs (8)	10
4. Improve scheduling*	7
5. Prove validity of tests	5
6. Improve introduction to guidance	4
7. Improve personal interest in counselee: take more interest (3), less interest (1)	4
8. Improve personnel: generally (2), no women (1)	3
9. Improve everything	2
10. Improve location	2
11. Others†	7

* At the time these veterans came to the center, veterans frequently had to wait five or six weeks between asking for guidance and the first appointment.

† Each of these was mentioned by only one veteran. They ranged from "advertise services offered" to "give veterans a letter of acceptance to school suggested."

ments according to the counselor involved did not indicate that the suggestions referred to one more than another.

The answers to item 12 covered a wide range. The median age for which counseling was considered desirable was 18 years.

Summary

The present study was undertaken to determine what attitudes toward the guidance process were held by 164 veterans who had received counseling at Stevens Institute in 1945 and 1946. In comparison with the data obtained from a non-veteran group which was counseled and later queried in a survey by the Psychological Corporation, the results were similar. The Stevens study suggests that in spite of limitations in time and money the veterans believed that they had profited from the VA program of educational and vocational guidance.

Received July 24, 1950.

Early publication.

Upper versus Lower Case Copy as a Factor in Typesetting Speed for Linotype Trainees *

Bernard Stern

State University of Iowa

The steadily increasing cost of labor in the newspaper industry has resulted in a continuing search for means and methods to bolster the productivity of newspaper employees. Especially is this true of mechanical employees. These workers, who belong to long established unions, are frequently paid more highly than editorial workers.

Thus, this writer was interested in noting in a book, *How to make type readable*, the following statement: "Material set in all capitals is read 12 per cent more slowly than material set in lower case. Reader preferences are overwhelmingly in favor of lower case."¹

Together with that claim by Paterson and Tinker was this statement: "It is apparent that all-capital text retards speed of reading to a striking degree. . . . Few typographical factors can be found which will retard reading to this extent."²

If these claims held true for newspaper linotype operators as well as for readers, it occurred to this writer that here lay a fertile means of boosting productivity. Such workers daily read and set in type millions of ems from wire copy which comes printed in all-capitals off thousands of teletype machines in newspaper offices throughout the country. In addition to this, these operators also set approximately as much type from reporters' typewritten copy which is usually set in lower case.

The problem, stating it more succinctly, is, in the light of Paterson and Tinker's finding, can linotype operators set more composition in a comparable amount of time from wire

copy (upper case) or from typewritten copy (lower case)?

If linotype operators could set an appreciably greater amount of copy from lower case copy rather than from upper case copy in a comparable amount of time, perhaps it would behoove the news wire services (Associated Press, United Press, International News Service) to make an adjustment in their teletype machines so that news stories could arrive in newspaper offices printed in lower case (caps and lower case). Or if this were not possible, it seemed likely that publishers could make a savings by employing typists to rewrite the all-capital wire copy into lower case.

On the other hand, if linotype operators could set more composition from upper case wire copy, would it not be worthwhile to have reporters turn in their stories typed in all-capitals?

An ideal means of testing this proposition with linotype operator trainees was available to the writer in the Newspaper Production Laboratory of the University's School of Journalism. Each semester the laboratory trains 12 to 15 student linotype operators.

It should be emphasized here that the operators participating in this experiment were beginners. At the time this study was conducted they were capable of turning out an average of one-half galley of composition an hour compared to experienced newspaper linotype operators who can set more than twice this amount per hour.

The findings set forth in this study, therefore, pertain solely to beginning linotype operators. No attempt is made to project the results to the work of experienced operators.

Procedure

This experiment was carried out after the student linotype operators had spent about twelve weeks in the Newspaper Production

* The writer would like to express his appreciation for the aid of Professor Clayton d'A. Gerken, Psychology Department; Professor Leslie G. Moeller, Marshall N. Heyman, William J. Morrison and Henry Africa, School of Journalism, State University of Iowa.

¹ Paterson, D. G. and Tinker, M. A. *How to make type readable*. New York: Harper & Brothers, 1940, p. 28 (obtainable from the authors, University of Minnesota).

² *Op. cit.*, p. 23.

Laboratory learning how to operate the linotype. Twelve students were used. This group was divided in two equal sections.

For this study, the author collected 130 news stories which had come over the Associated Press teletype in *The Daily Iowan*, student-published daily newspaper. These stories were cut into page lengths of $8\frac{1}{2}$ inches by 11 inches and were retyped in lower case. Every effort was made to get the same number of lines per page and the same number of words per line in the typewritten version as was in the wire (all-capital) story.

The 130 stories contained approximately 25,000 words and took up 110 pages. The various news stories varied in length from 60 words to 800 words and included the following varieties of news: international, national, crime and accidents, financial, reports of speeches and meetings, weather and sports. A breakdown in the sports category showed it included stories about baseball, golf, tennis, trapshooting, swimming and horse racing. Box scores and summaries of sports results usually set in agate type were omitted from the copy set by the linotype trainees. Each story was given a "slug" or tag to identify it.

Before the operators began work they were cautioned to set the stories exactly as they appeared on the copy. In the wire or all-capital version of the story the capital letter was underlined so the linotype operator would set that letter in a capital. Otherwise there was no difference between the two types of material.

Each of the twelve operators was given an opportunity to set the same stories from the two types of copy, upper case and lower case. If during the first week he worked on the all-capitals copy, the second week he set in type the same stories which had been retyped in lower case. If the operator had worked on lower case copy the first week, the ensuing week he was given the same stories to set from the upper case version.

A record was kept of the number of lines set daily by each operator from upper case or from lower case copy. At the end of each week these scores were totaled. The "*t*" test for the significance of the difference between the two groups of scores was then made.³

³ The "*t*" test employed was that for related measures, i.e., Theorem 4 in Lewis, D., *Quantitative methods in psychology*. Ann Arbor: Edwards Brothers, Inc., pp. 194-196.

Table 1
Operator's Weekly Average Speed, Lines per Minute and Weekly Average Errors, Errors per Line

Operator's Number (1)	Weekly Average Speed		Weekly Average Errors	
	Upper Case Copy (2)	Lower Case Copy (3)	Upper Case Copy (4)	Lower Case Copy (5)
2	1.50	1.29	.030	.045
3	1.28	1.74	.054	.021
4	1.50	1.44	.063	.043
5	1.13	1.20	.050	.034
6	2.10	1.76	.050	.062
7	1.25	1.54	.071	.024
8	1.35	1.57	.048	.040
9	1.32	1.40	.065	.190
10	1.48	1.47	.056	.058
11	1.44	1.35	.052	.059
12	1.25	1.53	.053	.038
14	2.16	1.97	.063	.054
Mean	1.48	1.52	.055	.056
S.D.	.74	.72	.028	.051
Mean Difference		.04		.001
S.E. Difference		.25		.043
<i>t</i>		.55		.078

Three linotype machines were available in the Newspaper Production Laboratory for the experiment. So that no one operator should have the advantage of working on a machine which might be easier to operate, the operators were rotated on the machines.

It was believed that as a corollary to determining whether linotype operators could set more composition from lower case or from upper case copy, it would be important to discover if the operators made more errors in setting upper case material than in setting lower case copy.

Before beginning the experiment, a pilot test was run. This was done with a different group of twelve linotype operators to see whether or not it would be possible to control the necessary factors. Another purpose was to find out if any "bugs" would arise to hamper the project.

The results of the pilot test were satisfactory and proved to be similar to those obtained in the experiment itself.

Results

Table 1, columns 2 and 3, shows the mean speed scores made by each of the twelve operators on the two kinds of copy.

The mean speed of the 12 operators on the upper case copy was 1.48 lines per minute, while on lower case copy, the mean speed was 1.52 lines per minute. The range of operators' performances varied from 1.13 to 2.16 lines per minute on upper case copy, and from 1.20 to 1.97 lines per minute on lower case copy.

Table 1, columns 4 and 5, shows the mean error scores made by each of the twelve operators on upper case and lower case copy. The mean errors of all operators on upper case copy was .055 errors per line compared to .056 errors per line on lower case copy. The range ran from .030 to .071 errors per line on upper case and from .021 to .190 on lower case copy.

At the bottom of Table 1, the mean difference, standard deviation, standard error of difference and the "*t*" test values are shown.

It was decided that a difference would not be considered significant unless the difference was within the five percent level. To be significant at the five percent level the "*t*" test value would have had to be 2.20.

No significant difference was shown in setting from upper case copy as compared to lower case copy. This was true for both speed and errors.

Conclusions drawn from this experiment are not presented as definitive proof that the findings apply to all operators. The students at the time they began this experiment had had twelve weeks of training during which period they had practiced an average of ten hours a week on the linotype. Only to this extent, can the findings of the experiment be applied generally and then they can be generalized for trainees only.

Although no attempt is made to project these findings, it may be helpful to know, simply as a frame of reference, what is expected of an experienced newspaper linotype operator. Three persons in the field with long experience, when consulted, agreed that an experienced operator could set between 1,600 and 1,700 lines of type a day and not make more than six errors per galley.

The speed of composition of an operator setting 1,600 lines would be 3.3 lines per minute. Figuring that there are 170 lines of type to a galley,⁴ an operator making six errors per galley would have an error of .035.

The average speed of all the linotype trainees on both types of copy (upper case and lower case) was 1.50 lines per minute. The average errors made by the twelve trainees on both types of copy was .0555. Thus it appears that the linotype operators participating in this experiment were, as a group, slightly less than half as fast and made nearly twice as many errors as an experienced newspaper linotype operator.

In concluding, it may be well to mention that linotype operators, by the very nature of their job, must read carefully every letter and symbol of the material they set in type. As a result their reading rate is slowed considerably since their job motivates them to stress accuracy in reading copy, rather than speed or comprehension. It is possible that this may account in large part for the lack of significant difference between the setting of the two different types of copy. Conversations with several

⁴ The figure 170 lines to the galley is based on the assumption the operator is setting 8-point type on an 8½-point slug and the slug is 12 ems wide.

linotype operators indicated that in setting ordinary material, they do not read it letter by letter unless such words are foreign to them, or such words are technical, complicated or names or the spelling is different than that to which they are accustomed.

As was stated previously, in setting copy experienced operators are required to set a certain amount of lines per day with a certain minimum of errors. Thus, in setting type there is a premium on both speed and accuracy. It may be added that it costs approximately fifteen cents per line to make a correction.

After the experiment was concluded, the participants were interviewed individually in an effort to discover which of the two types of copy they preferred, upper case or lower case.

Eight of the twelve unhesitatingly said they preferred lower case copy. They claimed it was easier to read and marking individual letters for capitalization sometimes was confusing. Two of the twelve were just as unhesitating in stating a preference for upper case. They said it was easier to read because the large cap letter was more easily perceived. The remaining two said they had no preference.

In matching the individual's record against his performance, it was ascertained that the attitude of the operator had little influence on what he did. In practically all of the cases, the operators sometimes would set the upper case copy faster one day and the lower case the next.

Received November 14, 1949.

Design Complexity as a Determiner of Visual Attention Among Artists and Non-Artists *

Walter A. Woods and James C. Boudreau

Pratt Institute, Brooklyn, New York

Observations by artists and art critics are frequently concerned with the need to train the artist to "see." Zadkin (8) speaks of teaching the pupil to see. Ensor (3) distinguishes between the practiced eye and the "more common eye." Roger Fry (4) comments: "Now this specialization of vision goes so far that ordinary people have almost no idea of what things really look like . . . the moment an artist who has looked at nature brings to them a clear report of something definitely seen by him, they are wildly indignant at its untruth to nature."

These observations are indicative of a rather widespread belief that the artist sees "better" or in a manner different from the non-artist. Preliminary attempts to investigate the visual patterns of artists and non-artists are briefly reported by Brandt (2) and Buswell (3). As yet no evidence has come to light which has a direct bearing on or which enables us to arrive at any general understanding of the problem.

The present experiment was designed to determine whether it was possible to discover differences in the "seeing" processes of artists and non-artists, and whether in fact, differences do exist. This report, which should be regarded as an exploratory experiment, suggests that it is possible to measure sensory differences in visual activity and that artists do differ, as a group, from non-artists.

Selection of a Visual Stimulus

The determination of the stimulus elements to be utilized stems from the hypothesis that artists will tend to respond differently from non-artists to designs of varying complexity; that the artist will be more sensitive and the non-artist less sensitive to complex designs. This hypothesis is derived theoretically from the long-established design concept of unity in

variety; that is, a satisfying work of art must at the same time be sufficiently unified so that attention will be centered on the design as a whole, yet sufficiently diverse so that interest is maintained. (See Graves (5) for a discussion of design principles.) Presumably the artist who is able to successfully execute a design which meets these criteria has command of a visual language of higher complexity than the non-artist who is unable to execute satisfying designs.

That artists differ from non-artists in their grasp of complexities of design units (and among themselves as artists) is suggested by the statement attributed to Picasso (6): "Cubism is no different from any other school of painting. The same principle and the same elements are common to all. The fact that for a long time cubism has not been understood and that even today there are people who cannot see anything in it means nothing. I do not read English, an English book is a blank book to me. This does not mean that the English language does not exist. . . ."

This statement illustrates one phase of our problem and at the same time raises another; what are the elements of design which are representative of "varying complexity"? Is color an element? Or form or space or subject matter? We are aware of the frequent criticism leveled at abstract art that it is "confusing" and meaningless; or the less frequent, but important criticism that the colors dominate the design and thus the meaning must be found in the "emotional" stimulus of the colors. It is generally understood that abstract art has some meaning (frequently believed to be a hidden meaning) and that this meaning (or the understanding of it) may depend upon the visual sophistication of the observer. According to some critics (and the above comment of Picasso bears this out) abstract designs do have symbolic meaning apart from their purely sensory qualities.

* The authors are indebted to Walter Civardi for photographic consultation and Eugen H. Petersen for aid in selecting and executing the designs.

Thus, an abstract design, as such, is not an adequate stimulus for our study since the grasp of symbolic content might enhance the attention value for the artistically sophisticated, regardless of sensory qualities.

The same criticism may be directed to color as a criterion. Colors are generally considered to have symbolic meaning (see Birren [1]) in addition to their stimulus value. Past experiences and associations may determine the attention value of any particular color or combination of colors, apart from such attention determining qualities of sensory appeal such as chromatic strength, reversal, dramatic quality.

Representational material (pictures of things) is equally unsatisfactory since the primary determiner of attention may be the symbols in themselves, apart from simplicity or complexity of the design. For example, a religious painting might very well attract strong attention from artist and non-artist alike if both were equally religious while the non-religious artist as well as the non-religious layman might reject the painting because it lacked satisfactory design qualities.

It became apparent that we were required to select, for our stimulus, designs which were neither representation nor abstract and which were monochromatic—which in fact were non-objective and which were black and white.

The Experiment

A well established technique for the measurement of differences in visual sensitivity (differences, actually, in eye movements) is that of Brandt (2). His report on the use of the Bidimensional camera suggested that this device and technique might be suitable for our experiment. It was further decided to utilize Brandt's plan of dividing the area into four stimulus areas. Design Charts A and B were executed in accordance with this plan, and with the above limiting qualifications: black and white and non-objective. The original designs were laid out on an area nine by fifteen inches, in the same proportions as shown on the accompanying charts. These designs were planned, as is apparent in Figure 1, so that they proceed in complexity from a single square to a cube with the interior exposed, and from a

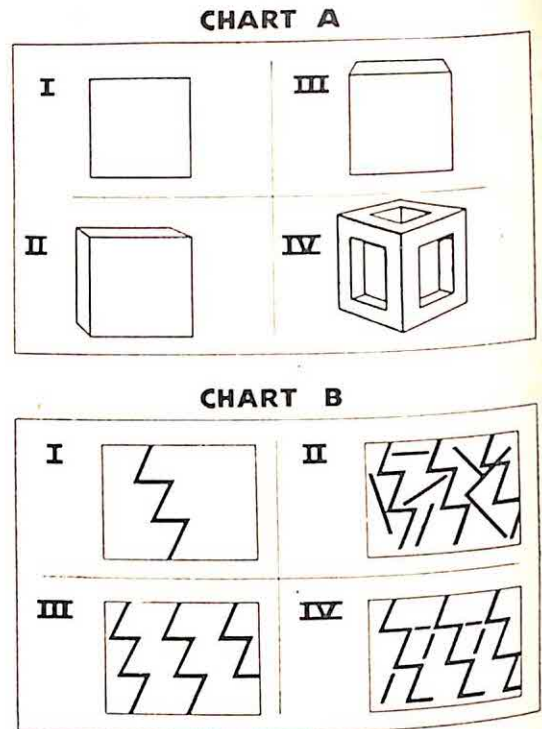


FIG. 1. Designs used in the experiment.

single broken line to three broken lines interrupted by diagonals.

Preliminary experiments were conducted to determine whether placement of the various units or areas would influence responses to the areas. Designs were rotated, following Brandt (2) in each of the four quadrants. It was apparent that position was not of significant influence insofar as individuals were concerned in the small sample used (eight individuals) so, in view of economy considerations (photographing and transcribing costs were prohibitive), it was decided to eliminate position as a control for the balance of the experiment.

The following groups of twelve subjects each were selected for the experiment: For design chart A: (1) Third year advertising design students; (2) First year (Foundation) art students; (3) Secondary school students age 15-17 taking art courses on Saturday; (4) Elementary school students age 12-14 taking art classes on Saturday; (5) Third year Electrical Engineering students selected for proficiency in mathematics; and (6) Third year Home Economics (foods) students. For design Chart B the following groups of twelve students each were used: (1) Third year advertising design

students; (2) First year (Foundation) art students; (3) Third year Electrical Engineering students selected for proficiency in mathematics; and (4) Secondary school students age 15-17 taking art classes on Saturday.

Although each group originally consisted of twelve students (except for the Foundation art group which consisted of thirty students) the final number of completed records was in all instances less than twelve. This situation came about because of faulty recording for the most part due to the use of release positive film. Actual numbers are indicated for each group in each of the following tables.

Employing the technique described by Brandt, the charts were placed in position on the rack of the camera and the subjects (observers) were asked to view the designs. Chart A was submitted first, to six groups; Chart B to four groups as noted above. The original intention was to use only Chart A, but after the experiment had commenced, it was decided to use Chart B as well as Chart A. The observer was allowed to look at the chart for nine seconds and then was instructed to close his eyes. Chart B was then placed in the rack and the observer instructed to open his eyes. He was allowed eight seconds to observe Chart B. Observers were given no instructions as to what to look for or how to observe the designs. They were simply instructed: "You will be shown some designs. Open your eyes when told to do so and close them when told to do so."

Responses of the observers were analyzed in terms of the following aspects of visual response: (1) Mean time spent by each group in observing each area of each chart; (2) Per cent of total time spent in viewing each area of each chart by each group; (3) Per cent of initial fixations made in each area by each group; (4) Per cent of total eye fixations made in each area by each group; and (5) Average number of fixations made by each group in each area. Analysis of variance was performed to determine variance ratios of mean time spent by each group in viewing each area, and to determine level of significance of these means.

Results of the experiment are shown in the following tables. It is indicated (Tables 1 and 4) that art students tend to spend more time in viewing the complex areas and less time in

Table 1

Per Cent of Time Spent in Viewing Each of Four Areas of Design Chart A by Members of the Following Student Groups: (1) Third Year Home Economics (Foods); (2) Students Age 12-14 Taking Saturday Art Classes; (3) Third Year Electrical Engineering Students; (4) Students Age 15-17 Taking Saturday Art Classes; (5) Students in Advertising Design; and (6) Foundation (First Year) Art Students

Group	N	Per Cent of Time in Each Area			
		Area 1	Area 2	Area 3	Area 4
Home Economics	6	20	24	26	30
Saturday Art					
Age 12-14	9	15	17	31	37
Electrical					
Engineering	8	20	22	22	35
Saturday Art					
Age 15-17	10	20	20	20	39
Advertising Design	9	12	20	22	46
Foundation Art	20	14	15	21	50

viewing the simple areas of the designs, whereas non-art students tend to distribute their time more evenly over the four areas. From Table 3 it will be seen that the between group variance exceeds the within group variance in

Table 2

Mean Time (in Seconds) Spent in Viewing Each of Four Areas of Design Chart A by Members of the Following Student Groups: (1) Third Year Home Economics (Foods); (2) Students Age 12-14 Taking Saturday Art Classes; (3) Third Year Electrical Engineering; (4) Students Age 15-17 Taking Saturday Art Classes; (5) Advertising Design; and (6) Foundation (First Year) Art Students

Group	N	Mean Time in Each Area				
		Area 1	Area 2	Area 3	Area 4	Total
Home Economics	6	1.84	2.16	3.34	2.59	2.25
Saturday Art						
Age 12-14	6	1.42	1.42	3.25	2.94	2.25
Electrical						
Engineering	6	2.00	1.94	2.09	3.00	2.25
Saturday Art						
Age 15-17	6	1.75	1.67	1.75	3.84	2.25
Advertising						
Design	6	.92	1.84	2.00	4.34	2.25
Foundation Art	6	1.17	1.09	1.94	4.84	2.25
Total	36	1.52	1.68	2.22	3.59	2.25

Table 3

Analysis of Variance of Mean Time Spent in Viewing Each of Four Areas of Chart A by Members of Six Student Groups (Based on Data in Table 2)

Source		df	Sum of Squares	Mean Square	F
Advertising Design	Total	23	213		
Between areas		3	161.5	59.83	
Among individuals		20	51.5	17.16	3.5
Foundation Art	Total	23	242		
Between areas		3	223.7	74.5	
Among individuals		20	8.3	.4	177.5**
Art Age 15-17	Total	23	152		
Between areas		3	80.3	26.8	
Among individuals		20	71.7	3.6	7.5*
Art Age 12-15	Total	23	202		
Between areas		3	68	22.7	
Among individuals		20	134	67	2.9
Electrical Engineering	Total	23	56		
Between areas		3	18.3	6.1	
Among individuals		20	37.7	1.9	3.2
Home Economics	Total	23	19		
Between areas		3	1.8	.6	
Among individuals		20	17.2	.9	1.4
Area I	Total	35	67		
Between groups		5	21.1	4.2	
Among individuals		30	45.9	1.5	2.8
Area II	Total	35	58.3		
Between groups		5	17.8	3.6	
Among individuals		30	40.5	1.4	2.6
Area III	Total	35	124.9		
Between groups		5	35.9	7.2	
Among individuals		30	89	3	2.4
Area IV	Total	35	241.6		
Between groups		5	83	16.6	
Among individuals		30	158.7	52.9	3.2
Total		143	884		
Between groups		23	550	23.9	
Within groups		120	334	2.8	
Areas		3	380.8	126.9	
Student groups		5	0.0	0.0	
Interaction		15	169.2	11.3	

Between groups F equals 8.5**
Areas F equals 11.2**
Groups F equals 0.0

the ratio of 23.9:2.8. These data, which indicate that a substantial proportion of variability is due to differences between the groups, are significant at the .001 level. A variance ratio of 22.1:3.7 for Table 6 is equally significant and supports the hypothesis that the various groups are (with exceptions noted by a more detailed analysis of variance and F

test found in Tables 3 and 6) representative of different populations.
In viewing Chart A (Table 1) first year art students spent the greatest per cent of time (fifty per cent of total time) in viewing area IV, advertising design students spent 46 per cent of their time in that area, and were followed in order by the 15-17 age group, the 12-14 age

group, engineering students and home economics students. It will be noted that the young art groups (age 12-14 and 15-17) both devote more time to the complex areas than do the older third year college engineers and home economics majors. It is indicated in this table that progressively more time is spent in viewing the complex areas by the art student and by the non-artist, but that the rates of progression are substantially greater for the art student. The same general pattern holds for Table 4 (per cent of time spent in viewing Chart B) except that here it is indicated that the engineer spends a somewhat greater proportion of time than the first year art student in viewing the more complex area III. The comparatively simple area I demands less attention by all except the secondary art student group. In this instance age (contrary to Buswell's findings) rather than artistic sophistication might be the determinant.

Tables 2 and 5 give mean time in seconds spent by each group in viewing Charts A and B and give the same information (in different form) as Tables 1 and 4. Tables 2 and 5 provide a basis for a determination of significance of differences and for analysis of variance.

Tables 7 and 8 indicate that initial eye fixations are primarily established in the upper areas (areas I and III), and are primarily directed to the right (observer's right). This is contrary to the findings of Brandt, and contrary to the generally accepted belief that initial eye fixations, determined by reading

Table 4

Per Cent of Time Spent in Viewing Each of Four Areas of Design Chart B by Members of the Following Student Groups: (1) Third Year Advertising Design; (2) Third Year Electrical Engineering; (3) Foundation (First Year) Art Students; and (4) Students Age 15-17 Taking Saturday Art Classes

Group	Per Cent of Time in Each Area				
	N	Area 1	Area 2	Area 3	Area 4
Advertising Design	8	11	43	16	30
Electrical Engineering	9	14	35	19	32
Foundation Art	14	14	32	20	33
Saturday Art Age 15-17	11	23	29.5	18	29.5

Table 5

Mean Time (in Seconds) Spent in Viewing Each of Four Areas of Design Chart B by Members of the Following Student Groups: (1) Third Year Advertising Design; (2) Third Year Electrical Engineering; (3) Foundation (First Year) Art Students; and (4) Students Age 15-17 Taking Saturday Art Classes

Group	Mean Time in Each Area					Total
	N	Area 1	Area 2	Area 3	Area 4	
Advertising Design	8	.87	3.5	1.25	2.38	2.0
Electrical Engineering	9	1.19	2.88	1.38	2.06	2.0
Foundation Art	14	1.87	2.25	1.38	2.5	2.0
Saturday Art Age 15-17	11	1.06	2.94	1.13	2.88	2.0
Total	42	1.25	2.89	1.28	2.58	2.0

habits, are most frequently directed to the observer's left. It will be noted that the non-art groups and the younger art groups tend to make more initial fixations to the left (area I). Two explanations are suggested: (1) The individual who is less sensitive visually will be dominated by reading habits while the visually sophisticated will be more attentive to visual forms and less dominated by reading habits or, (2) The artistic individual (or the visually sensitive individual) will be less likely to have developed strongly established reading habits and is therefore less dominated by habits which stem from reading. The latter view is in conformity with other data which suggests that a slight negative correlation exists between performance in art, as measured by grades in art courses, and verbal ability, as measured by the L score of the American Council on Education Psychological Examination (7).

Tables 9 and 10 demonstrate that the more artistically sophisticated groups tend to make fewer total fixations and tend to make a proportionately greater per cent of fixations in the more complex areas. The pattern of responses of the 12-14 age group indicates fewer fixations than is found for the 15-17 group.

Tables 3 and 6 give analysis of variance of the mean time spent in viewing each area by each group. Table 3 is based on means provided in Table 1 derived from responses to

Table 6

Analysis of Variance of Mean Time Spent in Viewing Each of Four Areas of Design Chart B by Members of Four Student Groups (Data from Table 2)

Source of Variation		df	Sum of Squares	Mean Squares	F
Advertising Design	Total	31	256		
Between areas		3	135	45	
Among individuals		28	121	4.32	10.4**
Electrical Engineering	Total	31	188		
Between areas		3	68.2	22.7	
Among individuals		28	119.8	4.28	5.3*
Foundation Art	Total	31	238		
Between areas		3	105.2	35.1	
Among individuals		28	132.8	4.74	7.4**
Art Age 15-17	Total	31	64		
Between areas		3	23	7.7	
Among individuals		28	41	1.46	5.3*
Area I	Total	31	56		
Between student groups		3	18.3	6.08	
Among individuals		28	37.6	1.3	4.7*
Area II	Total	31	52.1		
Between student groups		3	1.3	.43	
Among individuals		28	50.8	1.8	4.2
Area III	Total	31	185.5		
Between student groups		3	25.1	8.4	
Among individuals		28	160.4	5.7	1.5
Area IV	Total	31	170.2		
Between student groups		3	4.3	1.4	
Among individuals		28	165.9	5.9	4.2
Total		127	746		
Between groups		15	331.5	22.1	
Within groups		112	14.5	3.7	
Areas		3	246.4	82.4	
Student occupations		3	0	0	
Interaction		9	85	9.45	

Between groups F equals 6.**

Areas F equals 8.7*

Groups F equals 0.

Table 7

Per Cent of Initial Fixations Made in Each of Four Areas of Design Chart B by Members of Four Student Groups

	N	Area			
		1	2	3	4
Advertising Design	8	12.5	12.5	62.5	12.5
Foundation Art	14	7.1	7.1	78.7	7.1
Electrical Engineering	9	11.1	0.0	55.6	33.3
Saturday Art 15-17	11	36.4	0.0	45.4	18.1
Total	42	16.7	4.7	61.9	16.7

Table 8

Per Cent of Initial Fixations Made in Each of Four Areas of Design Chart A by Members of Six Student Groups

	N	Area			
		1	2	3	4
Advertising Design	8	25.0	0.0	37.5	37.5
Foundation Art	20	10.0	5.0	70.0	15.0
Saturday Art 15-17	11	36.3	0.0	45.6	18.7
Saturday Art 12-14	9	33.3	0.0	44.4	22.2
Electrical Engineering	9	44.4	11.2	44.4	0.0
Home Economics	6	50.0	0.0	33.3	16.7

Table 9

Per Cent of Eye Fixations Made in Each of Four Areas of Design Chart B by Members of Four Student Groups and Mean Number of Fixations for Each Group in Each Area

Group		Area				All Areas
		1	2	3	4	
Advertising Design	%	20.8	20.8	31.0	27.4	8.4
	mean N	1.5	1.5	2.25	2.0	
Electr. Engineer.	%	18.7	27.5	27.5	26.3	8.9
	mean N	1.67	2.44	2.44	2.33	
Foundation Art	%	17.8	21.4	32.2	28.6	10.0
	mean N	1.78	2.14	3.22	2.86	
Saturday Art 15-17	%	23.0	21.2	30.1	25.7	10.2
	mean N	2.36	2.18	3.09	2.63	

Chart A. Table 6 is based on analysis of variance of means in Table 3 and is derived from responses to Chart B. In general there is greater variation between area responses than among individuals in the groups. Home economics and engineering students tend to be most homogeneous in their visual patterns (less among individuals variance) and tend to view

Table 10

Per Cent of Eye Fixations Made in Each of Four Areas of Design Chart A by Members of Six Student Groups and Mean Number of Fixations for Each Group in Each Area

Group		Area				All Areas
		1	2	3	4	
Advertising Design	%	21.7	20.6	25.0	33.7	10.3
	mean N	2.2	2.1	2.1	2.6	
Foundation Art	%	19.6	21.3	26.4	32.4	11.7
	mean N	2.3	2.5	3.1	3.8	
Electr. Engineer.	%	27.6	20.0	22.8	29.6	11.6
	mean N	3.0	2.3	2.7	3.7	
Saturday Art 15-17	%	23.8	22.4	25.2	28.6	13.4
	mean N	3.2	3.0	3.4	3.8	
Saturday Art 12-14	%	22.9	20.2	26.6	30.3	12.1
	mean N	2.8	2.4	3.2	3.7	
Home Economics	%	22.0	26.8	22.0	29.2	13.7
	mean N	3.0	3.0	3.7	4.0	

the charts in a more homogeneous manner (less between areas variance) than do art students. Foundation art students are homogeneous as a group, but exhibit great variability in their response to areas. Only in the instance of the 12-14 age group do we find that variability due to differences among individuals within the group is greater than the variability due to differences in area complexity.

As to the variability which is brought about by differences between areas it may be noted that area IV (Table 3) draws the most variable response, both between the groups and among individuals within the groups, while the less complex areas I and III draw a much more uniform response. Variability due to differences among individuals is less than the variability due to differences between groups, except in the instance of area IV, wherein the among individuals variance is greater than the between groups variance.

In the assignment of total variance for all groups, between groups variance is substantially larger than within groups variance. The greater part of total variance must be assigned to differences between areas.

Summary

The present experiment attempted to arrive at some preliminary conclusions regarding the influence of design complexity in determining visual attraction or attention. It is indicated that art students tend to devote a greater proportion of their observation time to the more complex areas than do non-artists. Variance due to complexity of design is significant for the more sophisticated art groups, as are the differences in mean time spent in viewing the designs. However, significant differences are not found for the less sophisticated groups, indicating that sensitivity to design complexity is a developmental process which increases with age and with level of artistic sophistication. The general pattern of data support the hypotheses that: (1) differences in visual sensitivity or visual awareness and attention do exist between artists and non-artists and between age groups; and (2) art groups are more sensitive to or pay more attention to more complex design units than do non-artists

when the factors of color and objectivity have been eliminated from the designs.

Received November 4, 1949.

References

1. Birren, F. *The story of color*. Westport, Conn.: Crimson Press, 1941.
2. Brandt, H. F. *The psychology of seeing*. New York: Philosophical Library, 1945.
3. Buswell, G. T. *How people look at pictures*. Chicago: Univ. of Chicago Press, 1935.
4. Ensor, J. Selections from his writings. In Goldwater, R., and Treves, M. (Ed.), *Artists on art*. New York: Pantheon, 1947.
5. Fry, R. *Vision and design*. London: Wm. Clower and Sons, 1924.
6. Graves, M. *The art of color and design*. New York: McGraw-Hill, 1941.
7. Picasso, P. An interview. In Goldwater, R., and Treves, M. (Ed.), *Artists on art*. New York: Pantheon, 1947.
8. Woods, W. A. *Environmental influences in the development of persistence in space-form manipulation and their relation to artistic potential*. Ph.D. dissertation in process and yet unpublished, Columbia Univ.
9. Zadkin, O. *The poetic climate of art*. In Goldwater, R., and Treves, M., *Artists on art*. New York: Pantheon, 1947.

Verbal and Pictorial Questionnaires in Market Research *

Joseph Weitz

Carnegie Institute of Technology

The purpose of this study was to compare the results obtained from two different types of questionnaires commonly used in market research. The two questionnaires used in this study were a verbal and a pictorial questionnaire. Frequently in a single survey a questionnaire will be used which contains both verbal questions and a choice of pictorial items. The data obtained from these two types of questions are often treated similarly. It is hypothesized in the present study that different results may be obtained using these two techniques. If this is so, it does not seem warranted to treat the results in the same manner nor to evaluate equally the data obtained from these two sources.

Procedure

The subject to be studied in this survey is the design of the cooking range. This was chosen since it is generally present in some form in every household and hence people to be interviewed would be familiar with it. Two questionnaires were compared, one verbal and one pictorial, concerning the design of the cooking range. The questions on the verbal questionnaire were as follows:

- (1) Do you prefer the table top (low oven) or the high oven?
- (2) Do you prefer a window in your oven?
- (3) Do you prefer a high or low location for your broiler?
- (4) Do you prefer burner controls on the back vertical panel or the front panel?
- (5) Which of the following burner arrangements do you prefer?
 - (a) Two burners on each side with work space in the center
 - (b) Four burners on one side with work space on the other side
 - (c) Four burners staggered across the entire top
 - (d) Four burners across the back with work space across the front

* The author wishes to express his appreciation to Mr. David Ellies who did all of the interviewing and generally assisted throughout the entire project.

- (e) Two burners on each side with a built-in griddle in the center
- (6) Which of the following do you prefer?
 - (a) Oven in the center with storage space on both sides
 - (b) Oven on the right with storage space on the left
 - (c) Oven on the left with storage space on the right
 - (d) Double oven
 - (e) A high oven with storage space below
- (7) If you had your choice, what color would you choose for your stove?
- (8) Do you prefer to have toe space at the base of your stove?
- (9) Do you prefer a hinged door or a drawer type storage area?

The visual questionnaire was composed of a series of sketches involving the same discriminations which were asked for in the verbal questionnaire. For example, in question three—do you prefer the high or low location for your broiler?—in the pictorial questionnaire two drawings were made, one with a high broiler and one with a low, both stoves having the same basic design (see Figure 1). The person being interviewed was asked which of these stoves she would prefer. The same was done for all of the other questions.

For both groups other information was obtained, such as educational background, number of years the individuals had used the particular stove they now have, and what type of stove they were using at present. In this paper only the data pertinent to the original hypothesis will be presented.

The survey was conducted in the city of Pittsburgh. The sample consisted of 200 adult females. This total sample was divided into two groups of 100 each; one group received the verbal and one the pictorial questionnaire. In each sample of 100, 10% were from the A or highest socio-economic group; 30% from the B socio-economic group; 40% from the C socio-economic group and 20% from the D socio-economic group. In this way

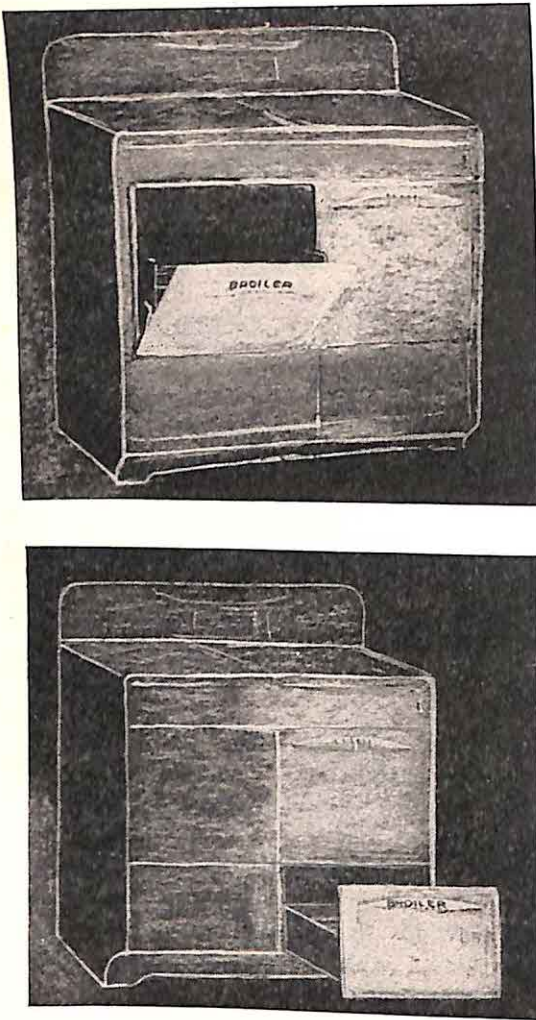


FIG. 1. Sketches used for question 3 in the Pictorial Questionnaire.

the two groups were matched on the basis of socio-economic background.¹

One interviewer was used for all 200 cases. This individual was an experienced interviewer. It was thought necessary to use only one interviewer so as to reduce that variable to a minimum. In both the verbal and pictorial questionnaires the interviewer returned to those addresses where no one was home at the first call. Further, in administering both the verbal and pictorial questionnaire the internal relationships of the sample were maintained during the study. That is, all of the A

group were not interviewed before the B, etc.; for example, 1 A was interviewed, then 3 B's, then 4 C's, then 2 D's, etc., so that no one group was completed before starting the next.

Results

The results of the study are shown in Table 1. It can be seen that in all cases, with the exception of question two, there was a significant difference at least at the 1% level as determined by the Chi-Square test.

On several items there was a complete reversal of the preferred response from the visual to the pictorial questionnaire. In question seven, concerning color preference, the colors other than white were grouped in one category and white in the other category, giving a two by two table rather than a five by two table, which would have been the case had all the colors been used (black, green, cream, blue and white).

Previously it was stated that even though no attempt was made to equate the two groups on the basis of educational background, they turned out to be quite similar. This can be seen from Figure 2. Since there would be fewer than five cases in some cells if Chi-Square were computed separately for each socio-economic group, all groups were combined to give the total educational level of each sample. From this combination of data the computed Chi-Square results in a value not significant at the 30% level. This then is an added check on the homogeneity of the two samples used and would lend weight to the assumption that any differences which exist are due primarily to the type of questionnaire used. It is of interest to note the change in educational background from the A group to the D group.

It can be seen from Table 1 that the results obtained from the pictorial and the verbal questionnaire cannot be considered as homogeneous. Since the sampling technique was identical for both samples and since there was evidence of homogeneity of the educational background of the two groups, it would seem evident that the differences observed are due to the differences in the questionnaire technique. If this is so, the original hypothesis is substantiated and one must differentially evaluate results obtained in these two ways.

¹ *Manual for research associates and interviewers*. New York: Market Research Division, The Psychological Corporation. Part 2, pp. 5-6.

Table 1
Preferences of Interviewees in Each Sample

Question	Choice	Number Preferring		Chi-Square Value	P
		Ver.	Pic.		
1. High or low oven	High	23	41	8.24	.010
	Low	77	59		
2. Window in oven	Yes	82	77	.764	.300 to .500
	No	18	23		
3. Broiler location	High	41	60	7.22	.010
	Low	59	40		
4. Burner controls	Front	9	27	10.98	.001
	Back	91	73		
5. Burner arrangement	2 each side	19	36	15.06	.010
	4 one side	57	30		
	4 across top	7	12		
	4 across back	11	11		
	2 each side griddle	6	11		
	Center	14	18		
6. Oven arrangement	Right	61	30	25.04	.001
	Left	6	5		
	Double	5	5		
	High	14	42		
	White	88	55		
7. Color	Not white	12	45	27.13	.001
	Yes	94	65		
8. Toe space	No	6	35	25.77	.001
	Yes	31	67		
9. Storage	Drawer	69	33	26.32	.001
	Hinge				

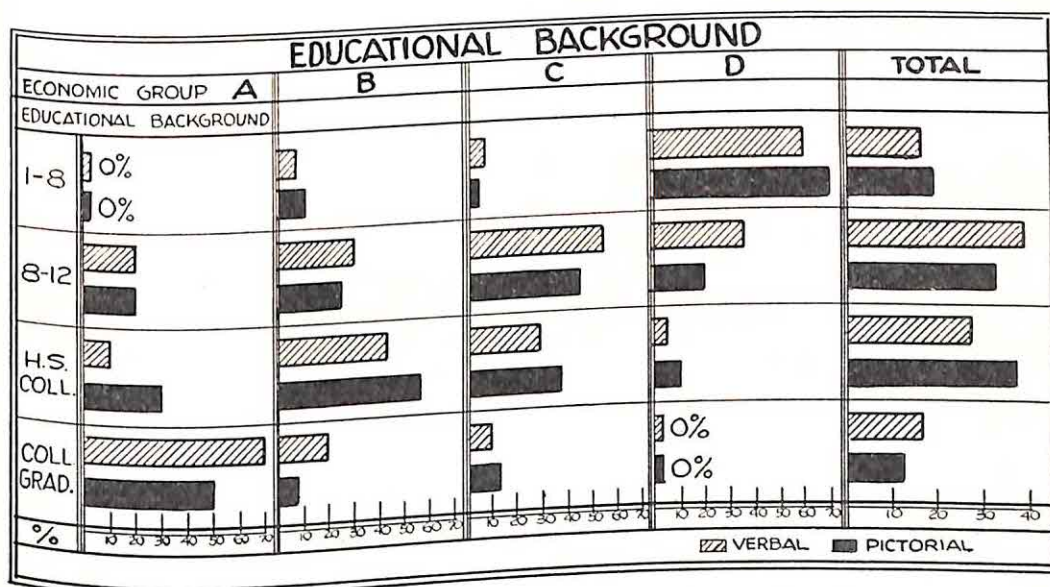


FIG. 2. Educational background of each socio-economic group in each sample.

The present study throws no light on the important problem of which of these two methods is more valid. That is, if a survey is to predict consumer behavior it should obviously be important to know which of these techniques, pictorial or verbal questionnaire, comes closer to actual buying behavior. Further research must be done in order to compare the usefulness of these two techniques with respect to their predictive value for consumer behavior.

It should be pointed out that it is possible that had other pictures been used there might have been different results; therefore the pictorial representation itself might be studied to determine the amount of variability obtained with various forms of visual presentation of questions.

Summary

Two different questionnaires were administered to two samples of 100 each. These samples were matched for socio-economic status and were homogeneous with respect to educational background. One group received a pictorial questionnaire, one group received a verbal questionnaire. Statistically significant differences were obtained between the two techniques. It is concluded from this study that one cannot use these two questionnaire techniques interchangeably and that the data obtained from these two methods should not be equally evaluated.

Received November 7, 1949.

An Exploratory Study of Linear Interpolation

Harry Kreiger Miller, Jr.

Lehigh University

The psychological problem of estimating the position of a point relative to two markers between which it falls is of vital importance in studies of dial and scale reading. Until recently there has been comparatively little scientific research published on this phase of interpolation. Work which has been carried out has indicated that there are three explanations for the errors which occur in such interpolations: (a) individual differences; (b) size of the interpolated interval and of the markers; and (c) biases.

Recognition that individual differences play an important part in this problem is indicated in a study on dial readings by Kappauf, Smith, and Bray (7) when they came to the conclusion that "subject differences and subject interactions appear to demand an analysis of the data subject by subject rather than analysis of group averages." In another study (9) on interpolation between circular scale markers, Leyzorek found that "individuals differ significantly in their abilities to perform this kind of visual interpolation."

That accuracy of interpolation varies according to the size of the interval has been observed in most of the studies on this subject, including the two just mentioned (7) and (9). Grether and Williams (5) concentrated on this factor in their experiment on accuracy of dial reading, and in another study (8) on dial size and graduation Kappauf and Smith found that frequency of dial reading errors was primarily a function of the size of the scale unit. The size of the markers also has some effect, as is mentioned by Bäckström (1) and some of the aforementioned.

Biases—or errors not randomly distributed—are the third category of causes of errors. Chapanis (3) found that whether the position of a point was underestimated or overestimated varied systematically with its location. Bartlett, Reed and Duvoison (2) and Leyzorek (9) found some evidence of the same type of biases. Reed and Bartlett (10) and Harriman

and Bartlett (6) studied these biases more specifically in experiments on concentric rings and positions along a short line, respectively. Bäckström (1) in studies on a limited number of subjects found that biases were of greater influence than random errors as a cause of discrepancies.

Purpose

The purpose of this research was to systematically study the accuracy of visual interpolation with five different interval sizes, each having interpolated positions 1 through 9, using twenty-one subjects of differing age and occupation.

Procedure

Figure 1 shows a pattern of the 2 mm. size. Six different patterns of problems were drawn up originally using an interval of 10 mm. Each pattern consisted of 54 problems, of which there were six each of the nine different positions (1, 2, 3, 4, 5, 6, 7, 8, 9) randomly arranged.

Direct prints were made of the original six patterns thus giving the 10 mm. interval. Photographic reductions were used for the 5 mm., 3 mm., 2 mm., and 1 mm. interval sizes. These were then placed on ordinary photographic mounting board.

A	1	2	3	4	5	6
	7	8	9	10	11	12
	13	14	15	16	17	18
	19	20	21	22	23	24
	25	26	27	28	29	30
	31	32	33	34	35	36
	37	38	39	40	41	42
	43	44	45	46	47	48
	49	50	51	52	53	54

FIG. 1. Sample pattern of 2 mm. size.

Slotted shields were used to permit viewing one horizontal row of problems at a time. Mimeograph answer sheets were provided.

Each subject was given one size of pattern at a time with the corresponding shield and answer sheets. He was asked to estimate in tenths the position of the inner line in relation to the two outside markers and to enter his judgment on the appropriate place on the answer sheet. He was told that the inner line always fell on an exact tenth (i.e., 1, 2, 3, 4, 5, 6, 7, 8, or 9). He was permitted to work at his own speed but was asked to be as accurate as possible with each interpolation. Use of the shield was optional, so that it was possible for him to compare problems against each other. Judgments were made at normal reading distance and no attempt was made to control lighting. No measurement of the subject's visual acuity was made.

When the subject finished one set he was given the opportunity of doing another size immediately or of resting his eyes before attempting the next group. The sets were not given in any systematic order, but rather the

subject was given his choice as to what order he desired to work on them. However, each individual completed all the problems of one size before being given the next set.

There were 520 judgments (6 patterns of 54 problems each) for each interval size, giving a grand total of 1,620 estimations by each subject.

Results

A tabulation of the total number of errors according to the size of the interval is made in Table 1. (Occasional mistakes due to reversal in recording—i.e., placing a 1 on the answer sheet for a 9, or a 2 for an 8—were not counted as errors of interpolation.)

The figures indicate that individual differences are of primary importance, since the totals extend from 3 errors for the most accurate subject to 538 errors for the least accurate.

Neither occupation nor sex is indicated as a major controlling factor. Of the four subjects making the fewest errors, two are males and two are females, one is a graduate engineering

Table 1
Number of Errors According to Size of Interval

Subject	Occupation	Age	Sex	Size of Interval					Total
				1 mm.	2 mm.	3 mm.	5 mm.	10 mm.	
RH	Grad. Eng. Stud.	25	m	0	0	1	2	0	3
FH	Housewife	21	f	0	2	0	2	0	4
GD	Grad. Educ. Stud.	21	f	3	0	0	0	1	4
LO	Undergrad. Eng. Stud.	27	m	1	2	0	3	1	7
RD	Undergrad. Eng. Stud.	25	m	3	1	11	0	0	15
FB	Housewife	23	f	7	1	4	4	1	17
HK	Grad. Psych. Stud.	26	m	10	0	4	1	6	21
RR	IBM Operator	21	m	6	4	16	2	0	28
GB	College Registrar	55	m	22	5	6	0	0	33
FF	Grad. Psych. Stud.	24	m	14	18	8	6	7	53
BG	7th Grade Stud.	12	m	35	13	16	15	4	83
EH	Undergrad. Psych. Stud.	23	m	6	6	11	26	36	85
ME	Office Clerk	18	f	33	9	21	17	11	91
JC	Housewife	23	f	38	16	43	10	16	123
WM	Undergrad. Bus. Stud.	25	m	42	47	25	33	23	170
HM	Accounting Clerk	50	m	58	31	26	24	48	187
DS	Housewife	50	f	101	31	29	18	20	199
ES	WAVE Officer	28	f	66	55	66	45	24	256
FL	7th Grade Stud.	12	m	116	43	36	32	37	264
BH	Office Clerk	19	f	122	48	74	45	67	356
WJ	Secretary	26	f	165	86	91	90	106	538

Table 2

Number of Errors According to Interpolated Position
 Note: L = Lower than true position; H = Higher than true position.

Subject	Interpolated Position														Totals		Grand Total				
	1		2		3		4		5		6		7					8		9	
	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H		L	H		
BG	0	1	1	3	4	0	31	0	3	1	1	25	0	8	0	2	3	0	43	40	83
EH	0	2	0	5	1	2	6	1	0	12	0	31	0	21	0	4	0	0	7	78	85
ME	0	5	2	5	5	5	27	6	6	0	5	5	6	3	2	6	3	0	56	35	91
JC	0	4	2	2	7	1	17	1	19	3	3	14	11	14	15	3	7	0	81	42	123
WM	0	17	0	5	14	0	57	1	0	3	14	23	3	19	1	0	4	0	102	68	170
HM	0	10	0	12	13	17	16	2	2	7	3	31	11	48	2	1	12	0	59	128	187
DS	0	2	3	8	25	3	38	3	10	7	7	42	4	37	3	4	3	0	93	106	199
ES	0	4	2	2	41	5	65	6	19	4	15	47	4	34	4	0	4	0	154	102	256
FL	0	0	6	3	25	2	56	1	21	18	1	62	1	51	0	15	2	0	112	152	264
BH	0	13	3	18	13	62	4	56	4	64	5	37	30	14	7	8	18	0	84	272	356
WJ	0	90	9	101	26	59	29	19	8	13	5	65	12	49	7	41	5	0	101	437	538

student, one is a housewife, one a graduate education student, and one a second year undergraduate engineering student. Of the four subjects making the most errors, three are females and one a male, all of different occupations.

Although most of the subjects are in the age-group of 18 through 28, the spread in total number of errors of the individuals outside this group lends credence to the inference that age is not a primary factor in accuracy.

In comparing the number of errors according to the size of the interval it was again found to be a matter of individual differences. Although the highest number of errors were made on the 1 mm. size by eleven of the subjects, there were two who had the greatest number of mistakes on the 2 mm. size, three who had the greatest inaccuracy on the 3 mm. size, two experienced the most difficulty on the 5 mm. intervals, and one who found the 10 mm. the most difficult for correct interpolation. In addition, one subject had an equal number of errors on the 2 mm. and 5 mm., and one found the 1 mm. and 3 mm. equally difficult. These data apparently point up the fact that there is no one optimum size for all subjects.

When the data were tabulated to show possible bias due to position, the first ten subjects in Table 1 showed so little error that there was no indication of bias. However, the last eleven subjects did show some trend and their results are included in Table 2. The smallest

number of errors is found on the 1 and 2, and the 8 and 9 positions, the extremes of the scale. The 5 position, the center of the scale, had comparatively few errors. The maximum number of misjudgments were made on the 4 and 6 positions. The 4's tended to be read as 3's, and the 6's as 7's, indicating a bias outward from the center. The 7's tended to be read as 8's, and the 3's as 2's.

Individual differences, however, are apparent. For example, subject BH overestimated the 3's, 4's, and 5's, and underestimated the 7's. Subject WJ made more mistakes on the 1's and 2's than on any of the others, whereas most subjects were comparatively accurate on these positions. Subject JC was least accurate on the 5's. Some subjects, for example EH, HM, BH, and WJ, made more overestimations, while others, as ME, JC, WM, and ES, had a greater number of underestimations.

Although the small sample of subjects does not permit statistical generalizations, the unusually high level of performance of some of the individuals is worthy of note. Four had less than $\frac{1}{2}$ of 1% errors, and the majority of subjects had less than 6% errors. No misjudgments were made on the 1 mm. size by the two most accurate subjects; this means that they discriminated differences of one-tenth of a millimeter.

Although some general trends are evident,

the individual is paramount above the size of the interval. For most subjects it became increasingly difficult to interpolate as the interval sizes decreased. Nevertheless, in some cases there was no decrease in accuracy and a few subjects did better at the smaller sizes.

Summary

In this study twenty-one subjects of differing age and occupation visually interpolated five interval sizes (1 mm., 2 mm., 3 mm., 5 mm., and 10 mm.), each size having 324 problems. The problems were randomly arranged in six different patterns, each pattern containing an equal number of the interpolated positions 1 through 9. The subjects were asked to be as accurate as possible and were given as much time as they needed to complete each interval size.

Eight of the subjects made less than thirty errors on the 1,670 problems. Results indicated that individual differences were of greater influence than interval size or biases due to position differences.

Received November 7, 1949.

References

1. Bäckström, H. Die Dezimalgleichung beim Ablesen von symmetrischen Skalen, ihre zeitliche Beständigkeit und ihre Abhängigkeit von der

- Ablesungsgeschwindigkeit, der Übung, u.a. *Z. Instrumkde*, 1940, 60, 231-237; 261-271.
2. Bartlett, N. R., Reed, J. D., and Duvoison, G. *Estimations of distance on polar coordinate plots as a function of the scale used*. Systems Research, The Johns Hopkins University. Report No. 166-I-44 (in press).
3. Chapanis, A. Accuracy of interpolation between scale markers as a function of scale interval number. *Amer. Psychologist*, 1947, 2, 346.
4. Chapanis, A., Garner, W. R., and Morgan, C. T. Instrument dials and legibility, and visual displays. Ch. V and VI in *Applied experimental psychology*. New York: Wiley, 1949.
5. Grether, W. F. and Williams, A. C., Jr. Speed and accuracy of dial readings as a function of dial diameter and angular separation of scale divisions. Ch. 7 in Fitts, P. M. (Ed.), *Psychological research on equipment design*. Washington: Gov't Printing Office, 1947, pp. 101-109.
6. Harriman, M. W. and Bartlett, N. R. Estimation of the relative position of a mark along a short line. (In press; private communication.)
7. Kappauf, W. E., Jr., Smith, W. M., and Bray, C. W. *A methodological study of dial reading*. Department of Psychology, Princeton University. Report No. 3, August 1947.
8. Kappauf, W. E., Jr. and Smith, W. M. A preliminary experiment on the effect of dial graduation and dial size on the speed and accuracy of dial readings. *Ann. N. Y. Acad. Sc.*, 1948 (in press).
9. Leyzorek, M. Accuracy of visual interpolation between circular scale markers as a function of the separation between markers. *J. exp. Psychol.*, 1949, 39, 270-279.
10. Reed, J. D. and Bartlett, N. R. Unpublished data, Systems Research, The Johns Hopkins Univ. (Private communication.)

Book Reviews

Libo, L. M. *Attitude prediction in labor relations—A test of understanding*. Studies in Industrial Relations, No. 10, Division of Industrial Relations, Graduate School of Business, Stanford University, 1948. Pp. 15. \$1.00.

The thesis of this monograph is that one first step to good industrial relations is the understanding on the part of labor and management of the points of view of the other party. Such understanding together with willingness to understand, ability to learn, and availability of the other party builds improved inter-personal and intergroup relationships. The extent of this understanding can be measured by comparing for each party its beliefs concerning the attitudes of the other party with the actual views of the other party. The study, then, essentially has two parts—the arguments concerning the major thesis and an illustrative attempt to measure the degree of understanding between labor and management.

It is apparent that this study holds that the keynote to improved labor-management relations is clarification. One might expect, therefore, that both concepts and terminology would be considered quite rigorously. While the author has striven mightily in this direction, in the reviewer's opinion he has not wholly succeeded in hitting his mark. Libo's concept of understanding appears to change from time to time. In some instances it seems to refer simply to the accuracy with which one party perceives the point of view of the other while in others it seems to imply acceptance of those points of view. Similarly in certain parts of the discussion understanding appears to be a necessary condition for improvement of labor-management relations and in other parts this position is denied. It is never quite clear whether understanding is concerned with knowledge of the attitudes, motivations, etc., of the other party or the prediction of the most likely behavior under given conditions.

As an illustration of the way in which Libo believes understanding can be measured, he administered some 24 opinion questions concerned with labor-management issues to a

group of 44 labor leaders and to a group of 33 industrial relations directors. The members of each group not only indicated their opinions but also those that they would expect of the other group. Through an analysis of the responses one is able not only to observe similarities and differences between the opinions of the two groups but also to observe the accuracy with which each group can predict—i.e., perceives—the opinions of the other.

On all but two of the 24 issues the two parties expressed opposite opinions. Management's predictions of the opinions of labor were found to be considerably more accurate than labor's predictions of the opinions of management. Generalizations of these findings are particularly hazardous in view of the fact that the two groups utilized do not have direct dealings with each other. The labor group consisted of representatives of the International Longshoremen's and Warehousemen's Union (CIO), and the management group were members of the California Personnel Management Association and each was from a different company.

On the positive side it can be said that the entire monograph is most stimulating. It is fruitful of ideas and should provoke further interesting discussion and research. Those interested in problems connected with such studies undoubtedly can profit from Libo's analysis of the problem of measuring understanding.

For those who may wish to refer to this monograph, it will be noted that the date is not cited. The reviewer has been informed that the date of publication is August 1948.

Edwin E. Ghiselli

*University of California,
Berkeley, California*

Van Der Lugt, M. J. A. *V. D. L. Adult psychomotor test series for the measurement of manual ability*. New York: New York University, 1948 (mimeographed); *V. D. L. Psychomotor test series for children for the measurement of manual ability*. New York: New York University, 1948 (mimeographed).

This series of ten relatively simple tests "for the measurement of manual ability" requires

only a small amount of test equipment and simple instructions for administration and scoring. The reliabilities for each test are said to be "over .90." Five basic components of speed, pressure, accuracy, motor memory and coordination are each represented by two tests, as follows: 1. Speed-prehension; 2. Speed-asyndesia; 3. Pressure-reproduction; 4. Pressure-control; 5. Accuracy-steadiness; 6. Accuracy-precision; 7. Motor memory-direction; 8. Motor memory-spatial; 9. Coordination-static; and 10. Coordination-dynamic.

Thus far it would seem that the battery fulfilled most of the requirements of useful measuring devices, since it samples most of the basic aspects (speed, etc.) of manual activities, is simple, objectively scored, and has a reasonable number of cases for its tentative norms.

However the basic problem of psychomotor testing has nearly always been one of validity, or "what the tests purport to measure." Ideally a psychomotor test battery should measure representative samples of each component of motor action, e.g., speed, precision, etc., so that the subscores could be used to describe or predict an individual's skill or aptitude for each component of any complex skill which he might be expected to develop.

Several coefficients of validity are cited, correlations with five point ratings by employers of skill on the job in small factories: $r = .73$ for women and .65 for men on the static coordination test, and .51 for women and .58 for men on the dynamic coordination test. Unfortunately the numbers of cases are described only as "small" and therefore cannot be evaluated as to their representativeness.

The only adequate tests of validity for these tests would be a series of correlations with a

variety of practical manual performances, each on groups of persons large enough to permit known statistical significance and with criteria of known reliability. Fine (as contrasted with gross) psychomotor tests have long been known to be quite specific in nature or related only within narrow group factors of other tests, and the validity of fine motor tests for the prediction of manual skills has been adequately demonstrated for only a few practical skills (air crew specialists and rifle marksmen).

In view of the above facts the reviewer may be pardoned for urging a very strong skepticism as to the appropriateness of naming the V.D.L. battery a psychomotor test series "for the measurement of manual ability." The reviewer's own research strongly suggests that various components such as speed of fine motor skills are not primarily determined by any stable biological characteristic of the individual so much as they are determined by his happening to hit upon qualitatively different *work methods* which are often quite specific to each skill studied (cf. Seashore, R. H. Theoretical and experimental analyses of fine motor skills. *Amer. J. Psychol.*, 1940, 53, 86-98).

It is to be hoped that Dr. Van Der Lugt will soon be able to furnish more definite evidence as to the validity or non-validity of the battery or its subtests for the prediction of a representative series of specific practical manual skills. Likewise the anticipated value of the battery for the appraisal of educational, clinical or developmental status of the individuals should be experimentally verified before we are justified in assuming this type of validity for the tests.

Robert H. Seashore

Northwestern University

New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor,
Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota

- Beginning experimental psychology.* S. Howard Bartley. New York: McGraw-Hill Book Co., Inc., 1950. Pp. 483. \$4.00.
- Teaching the child to read.* Guy L. Bond and Eva Bond Wagner. Revised edition. New York: The Macmillan Co., 1950. Pp. 467. \$3.75.
- A history of experimental psychology.* Second edition. Edwin G. Boring. New York: Appleton-Century-Crofts, Inc., 1950. Pp. 775. \$6.00.
- Developing men for controllership.* T. F. Bradshaw. Boston: Division of Research, Harvard Business School, 1950. Pp. 232. \$3.25.
- Selected readings in social psychology.* Steuart Hender-son Britt, Editor. New York: Rinehart and Co., Inc., 1950. Pp. 507. \$2.00.
- Recent experiments in psychology.* Second edition. Leland W. Crafts, Elsa E. Robinson, Theodore C. Schneirla, and Ralph W. Gilbert. New York: McGraw-Hill Book Co., Inc., 1950. Pp. 491. \$3.50.
- Rating employee and supervisory performance.* M. Joseph Dooher and Vivienne Marquis, Editors. New York: American Management Association, 1950. Pp. 192. \$3.75.
- Handbook of employee selection.* Roy M. Dorcus and Margaret H. Jones. New York: McGraw-Hill Book Co., Inc., 1950. Pp. 349. \$4.50.
- S. Weir Mitchell, novelist and physician.* Ernest Earnest. Philadelphia: University of Pennsylvania Press, 1950. Pp. 278. \$3.50.
- Be your real self.* David Harold Fink. New York: Simon and Schuster, Inc., 1950. Pp. 307. \$2.95.
- You and your health.* J. Roswell Gallagher. Chicago: Science Research Associates, Inc., 1950. Pp. 48. \$6.00.
- Studies in leadership: leadership and democratic action.* Alvin W. Gouldner, Editor. New York: Harper and Brothers, 1950. Pp. 768. \$5.00.
- The theory of mental tests.* Harold Gulliksen. New York: John Wiley and Sons, Inc., 1950. Pp. 462. \$6.00.
- General clinical counseling.* Milton E. Hahn and Malcolm S. MacLean. New York: McGraw-Hill Book Co., Inc., 1950. Pp. 373. \$3.50.
- Counseling the handicapped in the rehabilitation process.* Kenneth W. Hamilton. New York: The Ronald Press Co., 1950. Pp. 296. \$3.50.
- Chats with teachers about counseling.* S. A. Hamrin. Bloomington, Ill.: McKnight and McKnight Publishing Co., 1950. Pp. 224. \$3.00.
- How to use psychology for better advertising.* Melvin S. Hattwick. New York: Prentice-Hall Inc., 1950. Pp. 376. \$5.75.
- Speech therapy for the physically handicapped.* Sara Stinchfield Hawk. Stanford: Stanford University Press, 1950. \$4.00.
- Testing results in social casework.* J. McV. Hunt, Margaret Blenkner, and Leonard S. Kogan. New York: Family Service Association of America, 1950. Pp. 64. \$2.00.
- Measuring results in social casework.* J. McV. Hunt and Leonard S. Kogan. New York: Family Service Association of America, 1950. Pp. 79. \$1.50.
- Adolescent development.* Elizabeth B. Hurlock. New York: McGraw-Hill Book Co., Inc., 1950. Pp. 566. \$4.50.
- The principles of psychology.* William James. New York: Dover Publications, Inc., 1950. Pp. 700. \$7.50.
- Psychology in everyday living.* Ralph L. Johns. New York: Harper and Brothers, 1950. Pp. 564. \$3.50.
- A comparison of diagnostic and functional casework concepts.* Cora Kasius, Editor. New York: Family Service Association of America, 1950. Pp. 169. \$2.00.
- Tensions affecting international understanding.* Otto Klineberg. New York: Social Science Research Council, 1950. Pp. 227. Cloth, \$2.25; Paper, \$1.75.
- The growth and development of executives.* Myles L. Mace. Boston: Division of Research, Harvard Business School, 1950. Pp. 200. \$3.25.
- The meaning of anxiety.* Rollo May. New York: The Ronald Press Co., 1950. Pp. 376. \$4.50.
- Surveys, polls and samples.* Mildred B. Parten. New York: Harper and Brothers, 1950. Pp. 624. \$5.00.
- The criminality of women.* Otto Pollak. Philadelphia: University of Pennsylvania Press, 1950. Pp. 180. \$3.50.
- The Porteus Maze Test and intelligence.* Stanley D. Porteus. Palo Alto: Pacific Books, 1950. Pp. 194. \$4.00.
- Psychology, a biosocial study of behavior.* E. Terry Prothro and P. T. Teska. Boston: Ginn and Co., 1950. Pp. 546. \$3.75.
- Psychological problems in mental deficiency.* Seymour B. Sarason. New York: Harper and Brothers, 1950. Pp. 365. \$5.00.
- Religion and the cure of souls in Jung's psychology.* Hans Schaer. New York: Pantheon Books, Inc., 1950. Pp. 221. \$3.50.
- Occupational information.* Carroll L. Shartle. New York: Prentice-Hall, Inc., 1950. Pp. 339. \$3.50.
- Analytic group psychotherapy.* S. R. Slavson. New York: Columbia University Press, 1950. Pp. 275. \$3.50.

- The psychology of mental health.* Louis P. Thorpe. New York: The Ronald Press Co., 1950. Pp. 747. \$5.00.
- Educational psychology.* William Clark Trow. Boston: Houghton Mifflin Co., 1950.
- Management behavior and foreman attitude.* David N. Ulrich, Donald R. Booz, and Paul R. Lawrence. Boston: Division of Research, Harvard Business School, 1950. Pp. 56. \$.75.
- De profundis.* Oscar Wilde. New York: The Philosophical Library, 1950. Pp. 148. \$3.00.

The Concentric Organization Chart

C. G. Browne

Wayne University

The use of organization charts to illustrate personal and departmental relationships and the flow of authority and responsibility in any structured situation has long been a generally accepted procedure. Such charts perform the same functions with these non-quantified relationship variables as graphs do with quantified data. The organization chart also offers the same advantages of the graph, primarily the advantage of easier and quicker comprehension of the material under consideration.

It is the purpose of this article to present a type of organization chart diagramming, illustrated in Figure 2, which will be referred to as the concentric organization chart, and to indicate the advantages of this diagrammatic presentation of relationships over the more commonly used or traditional organization chart as illustrated in Figure 1. Both Figures 1 and 2 are based upon the organizational structure of a manufacturing company employing approximately 1,500 persons, all of the executives through the fourth echelon of the business being included in the charts illustrated.

It is proposed that, on the basis of the following considerations, the concentric organization chart is a more satisfactory representation of the material being presented than the traditional organization chart.

1. *Provides a better representation of the dynamics of personal relationships as they exist in an organizational structure.* The concentric organization chart represents dynamic, flowing relationships which proceed and are directed toward the center or inwardly, while at the same time there is a flow which proceeds and is directed from the center or outwardly. Characteristically, the functioning of any group or organization centers "around" individuals, not "below"

them. This is clearly presented in the circular plan in Figure 2, culminating in this case with the President and General Manager as the focal point for the activities of the other executives involved. Or it might be said that individuals in a group, whether executive or supervisory or not, are surrounded with contacts, influences, and relationships that stimulate them from all directions. The schematic representation of these factors is better diagrammed as a circular rather than a horizontal plane. These flowing, surrounding, focalizing relationships then take on the desirable aspects of a smoothly functioning organizational unit in that they represent schematically the cooperation, intercoordination, and mutual dependency which characterize successful group activities.

2. *Eliminates the "above and below" concept.* The importance of the principles of general semantics is gaining recognition in psychology, particularly where it can be shown that human relationships and reactions are causally related to words as conceptual symbols. Many words, phrases, and concepts take on or are promoted for a certain emotional significance which influences the individual. Most individuals readily recognize various levels of authority and responsibility, and will accept their personal position in relation to these various levels. However, an emotional aspect attaches itself to this recognition when it becomes a reflection on the status of the individual in the concept of above or below, higher or lower, superior or inferior, up or down, top or bottom. The concentric organization chart eliminates this emotional concept, and diagrams organizational structure in such a way that each individual is a supporting part of the structure, the entire structure being dependent on all of these

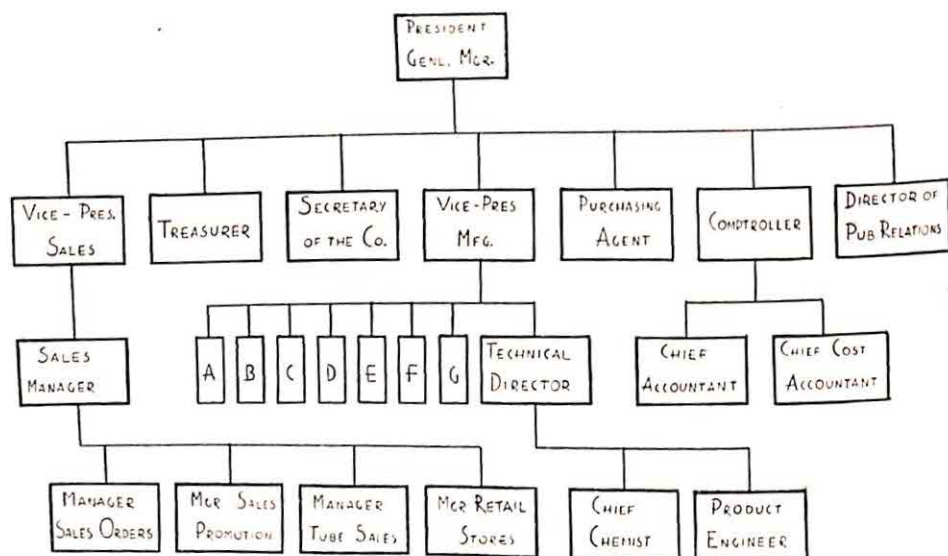


FIG. 1. Traditional type organization chart. Note: A. Mgr. Shipping; B. Mgr. Production Control; C. Mgr. Quality Control; D. Plant Engineer; E. Industrial Engineer; F. Personnel Director; and G. Foreman, Bicycle Tire Products.

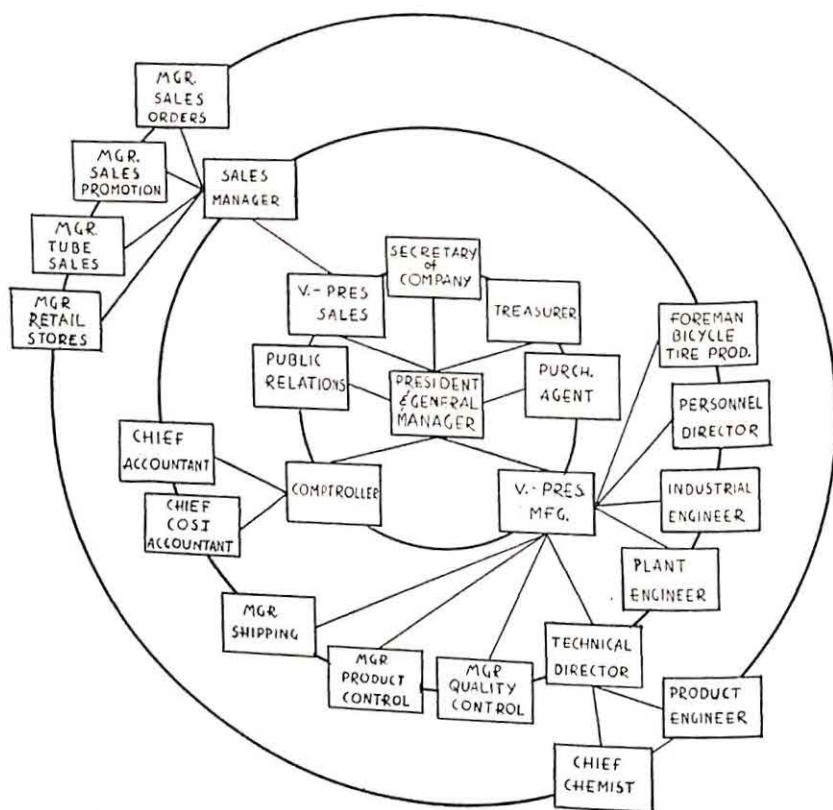


FIG. 2. Concentric organization chart. Note: Circles indicate echelon level; black connecting straight lines indicate flow of authority.

parts for its operation, since a break in any one of the concentric circles has an effect upon the entire organization and its functioning. Various echelons of authority are represented by the distance which any given circle is from the center or focal point.

3. *Presents an organization without loose ends.*

An organization is a self-contained unit within the sphere of its own operations. However, the traditional organization chart does not represent this, unless some limits are superimposed on the chart itself. Instead, the ends or the right and left of each echelon level represent a drop-off or vacuum of the organization's activities which is not characteristic of the functioning of and the relationships existing within the organization. In effect, the concentric organization chart brings these loose ends around into circular form in a diagram which represents a self-contained unit functioning on a continuum within its own boundaries as contained in the chart, all of which portray the actual functioning of the organization.

4. *Eliminates the upside-down organizational structure.* Since the concentric organization chart has neither top nor bottom, the interpretations of the relationships existing within the organization are not dependent upon the position of the diagram which represents them. Any organization chart is designed to present certain facts about a given organization at a given time. These facts or relationships do not depend on viewing the organization from any particular angle or any specific position, which is also the case for the concentric organization chart. However, with the traditional organization chart, the relationships which it intends to portray can be interpreted properly only when the chart is presented to the viewer in a certain position, that is, with the top at the top and the bottom at the bottom. Should this not be done, the organization would be upside down, with relationships existing inversely to the facts of the organizational structure.

5. *Simplifies designing and understanding.* After the facts of the organization chart have been determined, there is a difficult problem of the arrangement of these facts on the diagram.

This arrangement problem is particularly difficult on a traditional organization chart primarily because of the physical limitations of this type of chart. It is inevitable that the number of individuals will increase as the organizational set-up proceeds further and further from the focal point, or echelon one. That is, there is only one president, there may be four or five vice-presidents and other officers, 15 or 20 department heads, several hundred foremen, and thousands of non-supervisory personnel. However, on the traditional organization chart, the amount of space available for say the fifth or sixth echelon with 100 or more individuals is exactly the same space which is available for echelon one with one individual, the available space being a function of the width of the chart. This leads to many attempts and deviations by planners of organization charts to circumvent this problem in the best manner possible by changing the size of the box designating the individual (Figure 1, echelon 3), by raising or lowering the echelon level of some individuals in so far as the chart is concerned to an echelon where space is available (leading to confusion in following the relationships), by grouping individuals, or by any number of other devices.

The nature of the concentric organization chart automatically eliminates this primary problem of space in designing the chart. As illustrated in Figure 2, the necessity of including 11 executives on echelon three presents no problem since the third echelon circle is larger than the first and second echelon circles, where the number of executives is smaller. The same principle would follow in any expansion of the chart. This makes it possible to maintain a constant plan for showing the relationships between echelons and individuals, and for following these relationships through the system of connecting lines on the chart, which is not possible with the complicated network of lines and jumps to the left, right, up, or down which frequently create problems of interpretation and understanding with the traditional organization chart.

Received January 16, 1950.

Custom Made Systems of Job Evaluation

J. Stanley Gray

University of Georgia

Jobs have been evaluated since the beginning of employment but only in the last 25 years have any attempts been made to do this systematically. The term "job evaluation" now designates a *systematic*¹ attempt to weigh separately the worth of various elements or factors which constitute a job. This evaluation may be in money values (factor comparison), or point values (point systems), or some form of over-all classification. It is (or should be) preceded by job analysis, which also has been systematized in recent years. In other words, the trend in job evaluation and its concomitant activities has been toward systematization, with the assumption that more system will somehow result in greater validity.

This article presents one method of validating systems of job evaluation by elementary statistical procedure. Assuming that job evaluation is the determination of *relative* job values (i.e., the equitable distribution of a payroll—large or small) within an establishment, two other foundation theses are here presented. First, relative job values can be determined best by evaluating only those factors which are variables. If a factor exists in equal amount in all jobs being evaluated in an establishment, it is useless as a means of determining relative job values. This means that factors peculiar to an establishment must be used in evaluating jobs in that establishment. In other words *custom made* systems of job evaluation must be developed to fit each new job evaluation situation.

Second, the validity of any system of job evaluation (except factor comparison) can be determined by comparing its evaluation of key jobs with the wages already carried by those jobs. This means that when jobs are evaluated in points, the percentage of points as well as the percentage of wages for each key job can be calculated. It is then an easy matter to calculate the significance of the difference between the percentage of points

and the percentage of wages carried by each of the key jobs.

On the basis of these assumptions the writer has developed and validated custom made systems of job evaluation for establishments in both steel and textile industries and for the non-academic jobs in all 16 institutions of the University System in Georgia. In each case, a job evaluation system was developed to fit the peculiarities of the situation. It was then tested on key jobs and modified until it evaluated these properly. In the industrial situations, two lists of key jobs were used to double-test the validity of the system.

The procedure used may be described in seven steps:

1. All jobs to be evaluated are first carefully analyzed and described. Any thorough schedule of job analysis is appropriate but the one developed by the War Manpower Commission and reported in the *Training and Reference Manual for Job Analysis*, U. S. Printing Office, 1944 is recommended. Too often job evaluators study only those factors included in their ready made system and ignore other important aspects of a job. When a job is analyzed before the evaluation factors are known, the analysis is likely to be more thorough and more appropriate for other uses, such as employment, training, etc.

2. The next step is a careful analysis of the job analyses to *discover* the important variables which are to become the evaluation factors. Some of the standard factors—working conditions, experience, hazards, physical effort—are investigated to see if they differentiate jobs from each other. If they are fairly uniform from job to job they are *not* used as evaluation factors. The number of factors that are used depends on the number of differentiating characteristics. Frequency tables are made for each factor investigated.

3. The differentiating factors thus discovered are then assigned basic weights according to the estimated importance of each. The writer arbitrarily uses 100 as the total

¹ The word "scientific" is often erroneously used instead.

basic weight for all factors. The value of each factor is then the per cent of its importance in relation to other factors. This constitutes the value of the first degree. Higher degrees are arithmetically increased by multiplying this base value by the degree number. Thus factor A (perhaps working conditions) may carry a weight of 10 as a base (first degree) and 20, 30, 40, etc. for higher degrees. This means that if the first degrees of all factors total 100, the second degrees will total 200, the third 300, etc. Most point systems of job evaluation determine factor and degree weights in this way. Illustrative weights are given in Table 1.

A part of the weighting step in developing a custom made system of job evaluation, and yet a part that should come only after the weighting has been completed, is the development of a manual. Each factor must be carefully defined, as well as each degree under each factor. A job evaluation manual is intended to be used as a guide so that the rating of jobs will not be difficult. How much of a trait a job may have should be easily decided by referring to the manual. Job evaluation manuals should be written in the most simple yet the most exact language possible. They should be brief yet not ambiguous. The various key jobs can be added later as typical examples of the degree meaning of each factor.

4. The fourth step is to select 15 or 20 key or bench mark jobs, that is, jobs which already carry proper pay rates. These may be selected in various ways, but it is well to have an advisory committee composed of individuals who are intimately acquainted with the jobs to be evaluated.² Men who have worked in an establishment for years are better able to select key jobs than is a job evaluator who has only a short and incomplete knowledge of job duties and even less acquaintance with pay rates. *These key jobs should be representative of all the jobs to be evaluated.* If some department has no job that is properly paid, in the opinion of the committee, a typical job can be selected and a fair wage assigned to it even

² Both management and labor should be represented on this committee. It can be used later to advise regarding the evaluation of more difficult and complex jobs.

Table 1
Factor and Degree Weights for a System of
Job Evaluation

	Degrees				
	I	II	III	IV	V
1. Education	10	20	30	40	50
2. Experience	20	40	60	80	100
3. Physical Effort	10	20	30	40	50
4. Mental Effort	10	20	30	40	50
5. Working Conditions	10	20	30	40	50
6. Hazards	10	20	30	40	50
7. Responsibility for Safety	15	30	45	60	75
8. Tools, Materials, Equipment	15	30	45	60	75

though the actual wages are higher or lower. A key job with a hypothetical fair wage is just as good as one with an actual fair wage. With this system, jobs can be evaluated only as accurately as the rates on key jobs are accurate.

5. Obviously the next step is to evaluate key jobs, using the table of factors and weights as developed in Step 3. When the total point value of each key job is determined, a table of evaluation points and wages should be constructed. See Table 2, columns 2 and 4. Then the per cent of total evaluation points as well as the per cent of total wages should be calculated for each key job. See Table 2, columns 3 and 5. The difference between the percentage of total evaluation points and the percentage of total wages carried by a job indicates the accuracy of the job evaluation. See Table 2, column 6. The statistical significance of this percentage difference should then be calculated. See Table 2, column 7. If this difference is significant, it is evidence that the job is *not* accurately evaluated and the table of factor weights must be adjusted. Only when all key jobs are evaluated in such a way that no difference between percentage of total evaluation points and percentage of total wages is statistically significant can the evaluation system be called valid.

6. It is well to double check a job evaluation system, especially if the point values of factors and degrees had to be juggled repeatedly to fit the key jobs. This necessitates a second list of key jobs chosen with the same care as the

Table 2
Key Jobs with Wage and Point Values

Jobs	Evaluation Points	Per Cent of Total	Hourly Wage	Per Cent of Total	Difference	S.D. of Difference	Difference S.D. of Difference
Die Sinker A.	348	10.3	\$1.95	10.5	0.2	.71	0.28
Hammersmith (Steam)	321	9.5	1.75	9.4	0.1	.68	0.15
Millwright (Maintenance)	305	9.0	1.65	8.9	0.1	.66	0.15
Machinist (Maintenance)	286	8.4	1.55	8.3	0.1	.64	0.16
Shaper Operator	275	8.1	1.50	8.1	0.0	—	—
Crane Hitcher	260	7.7	1.45	7.8	0.1	.62	0.16
Grinder (Cutters)	241	7.1	1.30	7.0	0.1	.59	0.17
Inspector (Bench)	229	6.8	1.25	6.7	0.1	.58	0.17
Scale Blower (Forge)	214	6.3	1.18	6.4	0.1	.57	0.17
Furnace Unloader	204	6.0	1.10	5.9	0.1	.55	0.18
Roll Mill Helper	190	5.6	1.05	5.7	0.1	.54	0.18
Tester (Brinell)	179	5.3	1.00	5.4	0.1	.53	0.19
Drill Press Opr.	172	5.1	0.95	5.1	0.0	—	—
Straightener	164	4.8	0.90	4.8	0.0	—	—
Total	3388	100.0	\$18.58	100.0			

first. The jobs should be evaluated and a table constructed as described in Step 5. Again, if the differences between per cent of total points and per cent of total wages are great enough to be statistically significant, *the evaluation system is not yet valid*. Only when a job evaluation system justifies the wages of the key jobs can it be said to be valid enough to justify using it to evaluate other jobs which carry wage rates that may be too high, or too low, or just right.

7. The final step is to evaluate other jobs in the plant for which the system has been designed.³ This means the preparation of an evaluation sheet for each job, a summary sheet

showing the factor weights for all jobs, and a double entry graph showing the point values and present wage rates of all jobs. These results are then presented to the evaluation committee for their review and recommendations. These final steps are no different from those of any other job evaluation system.

Summary

Two innovations in job evaluation have been proposed in this paper. First, jobs are relatively evaluated by factors that represent differentiating characteristics between them. Ready made systems of job evaluation seldom fit because their factors do not differentiate between jobs. Valid job evaluation systems must fit the jobs being evaluated. Second, systems of job evaluation can be validated by computing the statistical significance of the differences between the evaluation of key jobs and the wage rates they carry. Key jobs are defined as those that already carry proper wage rates. When properly paid jobs are properly evaluated the evaluating system is then valid.

³ It must be remembered, however, that a custom made system of job evaluation is known to be appropriate for use with jobs typical of those represented in the key job sample. It may or may not fit jobs with different characteristics. If a group of jobs are found to be different from any one of the key jobs, it is then advisable to select the most typical one of them and add it to the other key jobs. This necessitates a repetition of Steps 4 and 5 and completely reworking the data tables. Often when jobs are greatly varied, it is necessary to develop more than one system of job evaluation. A system appropriate for evaluating production jobs is seldom appropriate for evaluating clerical jobs, or managerial jobs, or technical jobs. Different types of jobs are differentiated by different variables and should be evaluated by the use of appropriate factors.

Received January 6, 1949.

Rating Training and Experience

Laverne K. Burke and Erwin K. Taylor

*Personnel Research Section, AGO**

One of the most expensive, laborious and time-consuming chores in the examining process of any merit system is the evaluation of training and experience of applicants. It is also one of the least reliable and least objective aspects of the entire examining process.

As the importance of the position under consideration increases there is generally a concomitant increase in both the weight given to background factors and in the difficulty of evaluation. The task frequently requires the services of subject-matter experts beyond the staff of most examining agencies. Where the number of applicants is large, difficulty in securing an adequate number of qualified raters with time to devote to the task frequently delays completion of the register for a matter of months. Additional problems are introduced by the need to train raters to achieve some degree of uniformity in standards and to assure the applicants of bias-free evaluations.

A task of similar nature faced the Department of the Army when, after the war, a need was felt to re-evaluate Army officers. To accomplish the job the Officers Efficiency Evaluation Board was appointed to place all officers in a relative order of merit. An Army Personnel Records Board made available to the Evaluation Board brief but accurate and complete data on each officer's commissioned service prior to 1 January 1947.

The Army Personnel Records Board developed a form which has become known as the Statement of Service. On it the Board recorded each assignment, rank, rating, promotion, transfer, decoration, and disciplinary action for the period from January, 1937 through December, 1946. Pertinent remarks of superiors were also extracted from the efficiency reports and recorded on the form. Data covering all service prior to 1937 were more briefly summarized.

* The opinions expressed are those of the authors and do not necessarily represent official Dept. of the Army policy.

Complete statements of service were sent to the Officers Efficiency Evaluation Board. This group was charged with the responsibility of reviewing the record and arriving at an evaluation of the individual officer's value to the Army. The task of reviewing and rating some twenty to thirty thousand statements of service was indeed formidable. The Personnel Research Section of the Adjutant General's Office was called upon for assistance and was charged with the development of an objective clerical procedure whereby an initial ordering of officers could be accomplished. Members of

Table 1

Estimated Reliability of Rankings, by Branch
and by Sample
(In terms of average intercorrelation among 8 sets
of 10 ranks each)

Branch	Sample	
	A	B
Cavalry	.80	.84
Coast Artillery	.90	.91
Engineers	.89	.83
Field Artillery	.90	.74
Finance	.68	.82
Infantry 1	.94	.93
Infantry 2	.88	.74
Ordnance	.88	.69
Quartermaster	.90	.83
Signal Corps	.83	.81

the board planned to examine this listing and the records and make any necessary adjustments required by factors not taken into account by the first approximation.¹ The Officers Efficiency Evaluation Board's final evaluation, called the War Service Score, correlated .86 with the system presented here.

For purposes of analysis, a stratified sample of 200 lieutenant colonels was selected.

¹ The first approximation actually used was not the one here presented, but rather an improved system consisting of weighted composites of level of responsibility and efficiency reports.

Table 2

Correlation between Records Variables and Average Branch Rankings, Sample A

Note: All record variables, unless otherwise specified, are for the period from 1 Jan. 1937 through 31 Dec. 1946. Those marked by an asterisk are for the war period from 1 July 1943 through 30 June 1945.

Variable	Correlation Coefficient
1. Average efficiency rating	.81
*2. Average efficiency rating	.64
3. Average temporary grade held	.39
*4. Average temporary grade held	.56
5. Average: echelon of job X grade held	.59
*6. Average: echelon of job X grade held	.62
7. Per cent of war period spent overseas	.25
8. Age when Lt. Col. grade received	-.31
9. Present age	-.31
10. Number of times detailed to General Staff	.33
11. Total Army training received	.40
12. Highest Army training received	.44
13. Civilian education	.24
14. Years of enlisted service, not Regular Army	.09
15. Years of commissioned service, not Regular Army	-.09
16. Years of enlisted service, Regular Army	.02
17. Years of commissioned service, Regular Army	.02
18. Hospitalized, neurosis	.01
19. Hospitalized, physical	-.18
20. Highest grade reached	.69
21. Average temporary grade held (1942-1946)	.68
22. Highest efficiency rating	.15
23. Lowest efficiency rating	.69
24. Range of efficiency rating	-.68
25. Highest echelon assignment	.41
26. Total decorations received	.69
27. Highest decoration received	.67
28. Total number of assignments (1937-1941)	-.19
29. Total number of assignments (1942-1946)	-.13
30. Number of months unassigned, non-hospital	-.20
31. Number of months unassigned, hospital	-.13
32. Demotions	.63
33. Mention of special aptitudes by superiors	.23
34. Height	.01
35. Weight	-.06
36. Weight-height ratio	-.11

Twenty were selected at random from each of eight branches and 40 from infantry. The sample from each branch was sub-divided into

two groups, so that one stratified sample of 100 could be used for developing the evaluation system and the other comparable sample could be used for checking or cross validation.

For each branch subsample of ten cases the statements of service and special ranking forms were sent to the respective Career Management Branches. In each of these branches the ten officers in each subsample were *independently* ranked by eight officers who studied the statements of service carefully. To determine the reliability of these independent rankings, the average intercorrelation among rankers by branch was computed. These are presented in Table 1.

Careful analysis of the statement of service revealed 36 variables capable of quantification.

The remarks made by superiors when rating the officers were examined for the number and type of descriptive words and phrases used. However, consideration of the tendency of some superiors to put more descriptive remarks in their rating reports than others, even when rating the same officer, plus the length of time required to tally and check the number of different words used indicated that the analysis of the narrative material would not yield variables adequately reliable to justify their inclusion in the study.

The intercorrelations among the 36 variables from the statements of service and the average branch rankings were computed for 98 officers² in sample A. Table 2 presents the correlation of each of the 36 variables with the average rank. Table 3 presents the intercorrelations among the 36 variables.³ Applying the Wherry-Doolittle Test-Selection technique,⁴ a battery consisting of variables: (1) average efficiency report; (20) highest grade reached; (23) lowest efficiency report; and (27) highest decoration received, was developed.

The correlation of this battery with average

² Two officers in this sample had been prisoners of war and evaluation of their war service could not be made.

³ To reduce printing costs and to save space Table 3 has been deposited with ADI and may be ordered as Document 2749 from American Documentation Institute, 1719 N Street, N.W., Washington 6, D. C., retreating \$0.50 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$0.50 for photocopies (6 X 8 inches) readable without optical aid.

⁴ Stead, Shartle, & Associates. *Occupational Counseling Techniques*. New York: Am. Book Co., 1940. Pp. 245-252.

Table 4

Correlation between Branch Rankings and Composite of 4 Records Variables, by Branch and by Sample

Note: The combining weights were derived from the data of sample A and were used in combining the four variables for both samples.

Branch	Sample	
	A	B
Total	.90	.85
Cavalry	.93	.93
Coast Artillery	.90	.98
Engineers	.97	.92
Field Artillery	.98	.78
Finance	.77	.82
Infantry 1	.99	.82
Infantry 2	.94	.98
Ordnance	.93	.82
Quartermaster	.97	.93
Signal Corps	.79	.98

rank by branch is given in Table 4. To make sure that the high correlations in sample A (on which the test selection was based) were not merely a function of sampling error, the battery was also applied to the independent sample B. The correlation of battery scores with average ranking for sample B is also presented in Table 4.

It is interesting to note that there is better agreement between the battery score and the average ranking than there is between pairs of competent evaluators working from the same basic data. This fact argues strongly for the

use of a mechanical system of evaluating background items. Such a system not only results in considerable savings of time, money, and personnel but results in an evaluation that is consistently more reliable than the more laborious method of subjective evaluation.

Summary

As a possible means of evaluating Army officers' merit the Personnel Research Section developed a simple mechanical system of evaluation whereby four variables are objectively rated by reference to a set of specified weights. The system was cross-checked against a second group of officers and found to be effective. The evaluations made by the system agreed more closely with the consensus of groups of eight competent evaluators than pairs of evaluators agreed with each other in the subjective rating of the officers' entire records.

It is suggested that the technique employed could be of considerable service to merit systems. We estimate that it would be worthwhile doing when there are in excess of a thousand applicants to be rated. The system has the advantage that it will usually be possible to secure much more competent evaluators to make the trial rankings than could be secured to rate the entire applicant population. It is thus thought that application of the method described will not only save time and effort but will result in better evaluations as well.

Received February 1, 1950.

Measuring the Level of Abstraction

Rudolf Flesch

Dobbs Ferry, N. Y.

Students of communication agree that awareness of the level of abstraction is essential for full comprehension. For example, Perrin, in his college text, says (14): "For exact and reasonable communication it is highly important that a speaker or writer knows where in the range of meaning of abstract words his core of meaning falls and that he makes this clear to his listeners or readers." Semanticists are especially emphatic on this point. Johnson (10) writes: "The prime objective of general semantics is to make one conscious of abstracting." Hayakawa (8) writes: "Consciousness of abstracting is . . . a sign of adulthood."

Though the significance of the level of abstraction is widely recognized, no studies have been reported that attempt to estimate or measure it quantitatively. Discussions are usually limited to generalized distinctions between abstract and concrete words, illustrative examples, and such figurative devices as the semanticists' "abstraction ladder."

The present study is an attempt to approach this problem with the technique developed for the measurement of readability.

In recent years a number of statistical formulas to estimate readability (comprehension difficulty) have been developed (e.g., 3, 12), among them a four-part formula reported by the writer in 1948 (4, also 5). That formula was based on the following style elements: (1) average sentence length in words; (2) average word length in syllables; (3) average percentage of "personal words"; and (4) average percentage of "personal sentences."

To adapt the technique of readability measurement to the measurement of the level of abstraction, the writer utilized the fact that certain parts of speech are more frequent in abstract expression, while certain other parts of speech are more frequent in concrete expression. It is well known, for instance, that an abstract style contains relatively more descriptive adjectives, indefinite pronouns, and subordinating conjunctions, while a concrete style contains relatively more proper nouns,

limiting adjectives, finite verbs, personal pronouns, and coordinating conjunctions. Most of these statistical facts have never been used for the measurement of readability, but some of them have been the basis of such related studies as Boder's development of the adjective-verb quotient (1), Fries' analysis of the language of the Bible (7), LaBrant's study of subordination in children's writing (11) and Séchéhaye's study on the logical structure of the sentence (18). The present study is an attempt to combine and extend these statistical analyses of parts of speech and bring them in line with currently accepted grammatical classifications, as exemplified in the works of Curme (2) and Jespersen (9).

To sum up the hypothesis underlying the present study: Level of abstraction can be estimated by computing the ratio of certain parts of speech to certain other parts of speech in written expression. Since level of abstraction is a basic element in readability (comprehension difficulty), this ratio can be used as a measure of readability, either by itself or in combination with other elements.

Procedure

The criterion used in the writer's earlier readability formula (4) was McCall-Crabbs' *Standard test lessons in reading* (13). Since this criterion is keyed to children's grade levels, it seemed also a useful criterion for measuring the level of abstraction. It is generally agreed that children's mental growth corresponds to a growing ability to handle abstractions (see, for instance, Piaget, 15, 16, 17).

To make the prediction as accurate as possible, the following 42 of the 376 McCall-Crabbs' passages (containing poetry, directions, or arithmetical problems) were omitted from the count: Book II, Nos. 4, 42, 45, 77, 89, 92, 93, 94; Book III, Nos. 6, 11, 14, 15, 22, 32, 36, 41, 46, 51, 52, 63, 76, 78, 82, 84, 92, 94; Book IV, Nos. 8, 17, 31, 37, 43, 51, 75; Book V, Nos. 1, 21, 32, 52, 64, 79, 80, 83, 94.

To test the level of abstraction, the writer analyzed the generally accepted grammatical

categories of the parts of speech. It was found that most parts of speech contained certain categories that were statistically related to abstractness and certain others related to concreteness. In general, words related to abstractness are more indefinite, those related to concreteness more definite. The writer, therefore, chose the arbitrary label "definite words" for the words whose percentage was used to measure concreteness (level of abstraction).

The categories of "definite words," arrived at by a process of trial and error, were these (for more detailed definitions, see the section "How to Use the Formula" below):

- (1) The following nouns: Common and proper nouns with natural gender; common and proper nouns specifying time; nouns in the possessive case ending in 's or s'; and nouns modified by the limiting adjectives listed below.
- (2) The following limiting adjectives: Possessive adjectives; intensifying adjectives; numeral adjectives; the adjectives *what*, *this*, *these*, *that*, *those*, *each*, *same*, *both*; and *the* if modifying a noun not otherwise modified.
- (3) Finite verbs, except the verb *to be* used as copula.
- (4) Present participles ending in *ing* if used to form the progressive tense.
- (5) The following pronouns: Personal pronouns; reflexive pronouns; the relative pronouns *who*, *whose*, *whom*, *what*, *that*; interrogative pronouns; and the listed limiting adjectives when used as pronouns.
- (6) The following adverbs: *here*, *there*, *now*, *then*, *where*, *when*, *why*, *how*.
- (7) All interjections.
- (8) The words *yes* and *no*.

The percentage of these "definite words" in the test passages was correlated with the grade level of children who answered correctly one-half of the test questions.

To combine the new count of "definite words" with other tests of readability, intercorrelations were then computed. This yielded a multiple regression formula to predict readability, based on the percentage of "definite words" and the average word length in syllables. Following the procedure in the earlier study (5), this combined readability formula was then expressed in such a way that maxi-

mum readability had a value of 100 and minimum readability a value of 0.

Findings

The correlation between the percentage of "definite words" and the average grade level of children who could answer correctly one-half of the test questions was found to be $r = -.5541$. The corresponding regression formula is: $C_{50} = 8.7493 - .0852 dw$.

On the basis of this evidence, the percentage of "definite words" may be considered useful to form a rough estimate of the level of abstraction, ranging from 0 (fully abstract) to 100 (fully concrete).

The percentage of "definite words" may also be combined with average word length in

Table 1
Correlations, Means, Standard Deviations, and Regression Weights of Word Length and "Definite Words"

	dw	C_{50}	$\bar{X}$	s	β
wl	-.5681	.6908	135.1437	14.0030	.5552
dw		-.5541	37.9790	9.1720	-.2387

Table 2
Means and Standard Deviations of Two Criteria

	$\bar{X}$	s
C_{50}	5.5135	1.4096
C_{75}	7.1302	2.0454

syllables to form a measure of readability. The intercorrelations, means, standard deviations, and regressions weights found are shown in Tables 1 and 2. The following symbols were used: wl for word length (syllables per 100 words), dw for percentage of "definite words," C_{50} for the average grade of children who could answer one-half of the test questions correctly, and C_{75} for the average grade of children who could answer three-quarters of the test questions correctly.

The regression formula based on these correlations is: $R = 168.095 + .532dw - .811 wl$ (R represents the readability score).

Scores computed by this formula have a range from 0 to 100 for almost all samples

taken from ordinary prose. A score of 100 corresponds to the prediction that a child who has completed fourth grade will be able to answer correctly three-quarters of the test questions to be asked about the passage that is being rated; in other words, a score of 100 indicates reading matter that is understandable for persons who have completed fourth grade and are, in the language of the U. S. Census, barely "functionally literate." The range of 100 points was arrived at by multiplying the grade level prediction by 10, so that a point on the formula scale corresponds to one-tenth of a grade. However, this relationship holds true only up to about seventh grade; beyond that, the formula underrates grade level to an increasing degree. Finally, the formula—which predicted grade level and, therefore, difficulty—was "turned around" by reversing the signs to predict readability. (Before this transformation the formula read: $C_{75} = .0811wl - .0532dw - 1.8095$.) The multiple correlation coefficient of this formula is $R = .72$.

Since the correlations with the criterion C_{50} were higher than those with the criterion C_{75} , the multiple correlation with the criterion C_{50} was computed first; as a second step, the value so found was used to predict criterion C_{75} , in order to predict 75% comprehension rather than 50% comprehension.

Comment

The percentage of "definite words" appears to be a useful test in two ways:

First, it is a rough measure of the level of abstraction. As such, it may be used as a tool in semantic studies, critical reading, literary appreciation, translation, rating of advertising copy, and propaganda analysis. Step-by-step analysis of the level of abstraction in a given piece of discourse may also be helpful in logical analysis and the discovery of faults in reasoning.

Second, the new test is a measure of readability. As such it replaces the two tests of "human interest" (percentage of "personal words" and "personal sentences") that formed part B of the writer's earlier formula (5). Combined with the average word length in syllables, it gives a practical and comprehensive measure of readability.

It should be noted that this measure ap-

proaches the problem of readability in a new way, since it emphasizes the importance of the newspaperman's "Five W's" (Who? What? Where? When? Why?), the role of names and addresses, facts and figures, and the use of practical examples, illustrations, and anecdotes. Negatively, it points up the bad effects of too many adjectives, abstract nouns, and weak verb forms. It may, therefore, be expected to become a useful training device in the teaching of composition.

According to Shannon's mathematical theory of communication (19, 20), any noise hampering the transmission of a message is counteracted by redundancy in the message itself. By extension, any "semantic noise" (comprehension difficulty, ambiguity, etc.) is counteracted by "semantic redundancy" (repetition, padding, amplification, restatement, etc.). In other words, information on a high level of abstraction, which is difficult to grasp, is communicated more effectively to the extent that restatements on a more concrete level are added. This, of course, corresponds to the use of illustrations, practical applications, examples, parables, and the like in teaching. The new measure of abstraction may possibly be used in conjunction with the mathematical theory of communication to estimate the number and level of concrete illustrations necessary to convey a given piece of abstract information to a given audience.

This general principle has long been recognized in practice and there is a marked tendency in modern writing to alternate between abstract generalizations and concrete examples. Because of that, the percentage of "definite words" often fluctuates between extremes within a single piece of writing. This may make more or longer samples necessary in order to arrive at a reliable estimate.

The correlation of the new factor to comprehension (.55) is higher than that of average sentence length to comprehension (.52), as reported in the writer's earlier paper (5). In other words, concreteness appears to be somewhat more important for readability than the use of short sentences.

An amusing illustration of the new measure is furnished by Hayakawa (8), who explains his "abstraction ladder" by way of four statements on different levels of abstraction. Ap-

plication of the new measure to these examples shows the following:

	Ratio of "definite words"	Percent- age of "definite words"
<i>Mrs. Plotz makes good potato</i> <i>pancakes.</i>	3 in 6	50
<i>Mrs. Plotz is a good cook.</i>	2 in 6	33
<i>Chicago women are good cooks.</i>	1 in 5	20
<i>The culinary art has reached a</i> <i>high state in America.</i>	1 in 10	10

How to Use the Formula

For practical application, the directions for using the formula may be stated as follows:

To measure the level of abstraction of a given piece of writing, go through the following steps:

Step 1. Unless you want to test a whole piece of writing, take samples. Take enough samples to make a fair test. Don't try to pick "good" or "typical" samples. Go by a strictly numerical scheme. For instance, take every third paragraph or every other page. Each sample should start at the beginning of a paragraph.

Step 2. Count the words in your piece of writing or, if you are using samples, take each sample and count each word in it up to 100. Count contractions and hyphenated words as one word. Count as words numbers or letters separated by space.

Step 3. Count the following "definite words." Count each of these words only once. Count as words all units separated by white space. (In the examples, "definite words" are italicized.)

(1) Count all names of people—that is, proper nouns with natural gender, either masculine or feminine. Count all words that are part of the name, including titles, etc., used as part of the name. Example: *President Harry S. Truman* (count 4).

Count names of people used as adjectives to modify natural-gender nouns (e.g., the *Smith brothers*, the *Dolly sisters*), but do not count personal names used as adjectives to modify nouns without natural gender (e.g., the Ford Motor Company, the Washington Monument).

(2) Count all common nouns that have natural gender, either masculine or feminine. Examples: *father*, *mother*, *iceman*, *actress*. Count these words also when used as adjectives to indicate the gender of another noun (e.g., *woman doctor*, *bull elephant*, *girl athlete*), but do

not count them when they do not indicate gender (e.g., fellow workers).

Do not count common-gender nouns like teacher, doctor, employee, spouse.

(3) Count all nouns that indicate a specific time on the clock or the calendar; e.g., the names of the months, the seasons, the days of the week, the words *morning*, *noon*, *afternoon*, *evening*, *night*, *day* (in the sense of daytime), *sunup*, *sundown*, *today*, *tonight*, *yesterday*, *tomorrow*, and the words *breakfast*, *lunch*, *dinner*, *supper* when used to indicate time. Count these words also when used as adjectives to specify time (e.g., *December day*, *fall season*, *lunch hour*), but do not count them when used as adjectives without relation to time (e.g., Thursday club, dinner companion, Sunday suit).

(4) Count all numeral adjectives and all nouns modified by numeral adjectives. Count also the words *first*, *next*, *last*, and such words as *double*, *pair*, *half*, *triple*. Examples: *27 words*. *54 percent*. *Six letters*. The *next day*. The *last moment*. But don't count nouns that are not directly modified by numerals, e.g., *thousands of people*, a *ten-year-old* house.

Count the word *one* in the sense of "single," but not when used as an indefinite pronoun or as part of "no one," "any one," or "some one." Examples: Count *one* in "*one fine morning*" and in "*the pretty one*," but not in "*one has one's doubts*."

Count the word *once* in the sense of "a single time," but not in the phrase "at once." Examples: *Once upon a time*. *He succeeded once*.

Count the words *other* and *another* only in the sense of "second" or "one more."

(5) Count all finite verb forms—that is, verbs in the first, second, or third person and the present, past, or future tense. In verb forms with auxiliary verbs, count the auxiliary rather than the main verb. Examples: *He came* and *went*. *We have considered*. *You should have declined*. *Let us pack up and go*. *It was debated and voted down*.

Exception: Do not count the verb "to be" when used as copula, that is, as a link between subject and predicate. Do not count "to be" as copula in any form, with or without an auxiliary verb. Examples: *I was sick*. *You should have been there*. *It might be fun*. *He ought to have been careful*.

(6) Count all present participles (-ing) when used as part of the progressive tense (to be -ing). Examples: *He was running*. *I am going to look*. *You should have been working*.

(7) Count all personal pronouns except indefinite "it," and all reflexive pronouns, formed with -self or -selves. Examples: *It is a fact that you and I are not related*. *It was Mary herself*.

(8) Count the words *here*, *there*, *then*, and

now (except indefinite "there" in "there is," "there are," etc.).

(9) Count the words *who*, *whom*, *when*, *where*, *why*, and *how*.

(10) Count the words *this*, *that*, *these*, and *those*; *each*, *same*, and *both*; and nouns modified by them.

(11) Count the words *what* and *which* (interrogative) and nouns modified by them, but not the word "which" when used as a relative pronoun. Examples: *Which way are you going?* But: *The car which I bought.*

(12) Count all possessive pronouns (*my*, *your*, *his*, *her*, *its*, *our*, *their*, etc.), all nouns in the possessive case ending in 's or s', the word *whose*, and all nouns modified by these possessives. Examples: *Our modern civilization, its recent development, whose business, journey's end.*

But do not count possessive cases of pronouns not otherwise counted, e.g., "one's ideas," "someone else's hat."

(13) Count the word *that* when used as a relative pronoun, but not when used as a conjunction. Examples: Remember that *you* are sick. It was the humidity *that* did it.

(14) Count the words *yes* and *no* (used as answer).

(15) Count all interjections.

(16) Count the definite article *the* and the noun modified by it, but only if that noun is a single word not otherwise modified. Examples: *I missed the bus. We called the doctor. It was the truth.* But: *I missed the green bus. We called the eye specialist. It was the truth, plain and simple. They gave me the room I had last time. It was the beginning of the end. We live at the end of the village, in the yellow house at the turn of the road.*

Do not count the word "the" when modifying a noun that is to be counted under one of the other definitions. Examples: *He came in the afternoon. The only one there was the boy.*

Do not count the word "the" when modifying adjectives or noun-adjectives, particularly proper noun-adjectives referring to nationality, race, etc. Examples: *Their team was the best. We sat in the dark. What's the good of it? The Scotch are thrifty. The Negro was arrested.*

Step 4. The number of "definite words" in your 100-word sample (or the average percentage of "definite words" in all your samples or the whole piece of writing tested) indicates the level of abstraction. The typical relation between the percentage of "definite words" and level of abstraction is shown in Table 3.

To test the readability of a piece of writing, go through the following additional steps:

Step 5. Count the syllables in your 100-word samples or, if you are testing a whole piece of writing, compute the number of syllables per 100 words. If in doubt about syllabi-

Table 3

Typical Percentages of "Definite Words"

Level of Abstraction	Percentage of "Definite Words"
Highly abstract	Up to 20
Fairly abstract	20 to 30
Fairly concrete	30 to 45
Highly concrete	Over 45

cation rules, use any good dictionary. Count the number of syllables in symbols and figures according to the way they are normally read aloud, e.g., two for \$ ("dollars") and four for 1918 ("nineteen-eighteen"). If a passage contains several or lengthy figures, your estimate will be more accurate if you don't include these figures in your syllable count. In a 100-word sample, be sure to add instead a corresponding number of words in your syllable count. To save time, count all syllables except the first in all words of more than one syllable and add the total to the number of words tested. It is also helpful to "read silently aloud" while counting.

Step 6. Find your readability score by inserting the number of "definite words" per hundred words (*dw*) and the number of syllables per hundred words (word length, *wl*) in the following formula:

$$R = 168.095 + .532 dw - .811 wl.$$

The readability score will put your piece of writing on a scale between 0 (practically unreadable) and 100 (easy for any literate person). Readability scores will tend to follow the pattern shown in Table 4.

Table 4

Pattern of Readability Scores

Readability Score	Description of Style	Syllables per 100 Words	"Definite Words" per 100 Words
0 to 30	Very difficult	192	5
30 to 50	Difficult	167	14
50 to 60	Fairly difficult	155	22
60 to 70	Standard	147	29
70 to 80	Fairly easy	139	36
80 to 90	Easy	131	43
90 to 100	Very easy	123	50

Or you may want to test a piece of writing separately for concreteness and "reading ease." If so, do this after **Step 5**:

Alternative Step 6: Figure the average sentence length in words for your piece of writing or, if you are using samples, for all your samples combined. In a 100-word sample, find the sentence that ends nearest to the 100-word mark—that might be at the 94th word or the 109th word. Count the sentences up to that point and divide the number of words in those sentences by the number of sentences. In counting sentences, follow the units of thought rather than the punctuation: usually sentences are marked off by periods; but sometimes they are marked off by colons or semicolons—like these. But don't break up sentences that are joined by conjunctions like *and* or *but*.

Alternative Step 7: Find your "reading ease" score by inserting the number of syllables per 100 words (word length, *wl*) and the average sentence length (*sl*) in the following formula:

$$\text{R. E. ("reading ease")} \\ = 206.835 - .846 \text{ } wl - 1.015 \text{ } sl.$$

The "reading ease" score will put your piece of writing on a scale between 0 (practically unreadable) and 100 (easy for any literate person).

Sample Application

To show the application of the new test, three passages dealing with the subject of truth will be used. Passage A, exemplifying a highly concrete style, is Test Lesson No. 37 from Book III of the *Standard Test Lessons in Reading* (13) that were used as criterion. Passage B was taken from *Courts on Trial* by Jerome Frank (6). It is an example of writing on the middle level of abstraction, illustrating the more abstract generalizations in the first paragraph by more concrete examples in the second. Passage C was taken from *The Next Development in Man* by L. L. Whyte (21) and shows a highly abstract, philosophical discussion of truth.

"Definite words" in the three passages are italicized.

Passage A:

The children were telling about their Christmas vacations.

"We went to Kansas," said Jack. "One day when we were skating on the lake some of the boys cut a hole in the ice, struck a match and a fire blazed right up out of the hole for two or three minutes."

"Oh, oh!" said all the others, "that couldn't be true. Water doesn't burn."

"But it is true," said Jack. "I saw it."

They turned to the teacher to see what she would say and she explained this very strange happening. It seems there are natural gas wells under the lake which send the gas bubbling up through the water where it is caught in large pockets under the ice.

"So you see," said the teacher, "when a hole is cut the escaping gas will burn if lighted."

Passage B:

No means *then*, have as yet been discovered, or are likely to be discovered, for ascertaining whether or to *what extent* the belief of the trial judge about the facts of a case *corresponds* to the objective facts as they actually occurred, when the witnesses disagree, and when some of the oral testimony, taken as true, will support the judge's conclusion. In other words, in such a case there is no objective measure of the accuracy of a judge's finding of the facts. There exists no yardstick for that purpose.

In a "contested" law suit, therefore, with the witnesses in disagreement, usually no one can adequately criticize the trial judge's fact-finding. If, at the end of the trial, the trial judge says that *Jones hit Smith*, or that *Mrs. Moriarity called Mrs. Flannagan a liar*, or that *old widow Robinson was insane when she made her will*, or that *Wriggle used fraud in inducing Simple to sign a contract*—the judge's word goes. And the same would be true if, in most of those instances, the trial judge had found exactly the opposite to be the facts.

Table 5
Analysis of Three Sample Passages

	Percentage of "Definite Words"	Syllables per 100 Words	Readability Score
Passage A	53 (highly concrete)	121	98 (very easy)
Passage B:		151	58 (fairly difficult)
First paragraph	24 (fairly abstract)	152	65 (standard)
Second paragraph	37 (fairly concrete)	151	62 (standard)
Both paragraphs	31 (fairly concrete)	159	48 (difficult)
Passage C	17 (highly abstract)		

Passage C:

Truth is thought which *conforms* to the form of the whole. Conformity to the whole is the *criterion*. The unitary truth is *that* which conforms to the whole process of which it is a part. *The truth* is a form embedded in the whole complex of processes in the human organism and *its environment*, symbolizing and organizing them. A particular truth *may* not represent the entire structure of a situation but only *those aspects* which are relevant to thought at a given stage in *its development*. *The truth* is a system of symbols *whose structure conforms* to the whole pattern of feeling, thought, and action, and *integrates* all the processes which *link* the reception of stimuli and the molding of the ultimate responses. *This* is not a pragmatic criterion, since unitary truth *does* not merely serve special needs but *unites* the whole system in a conviction which is at once emotional, intellectual, and practical.

A comparative analysis of the three passages is given in Table 5.

Considerable experience with practical application of the new test has shown that an untrained person can acquire reasonable familiarity with this test after ten to twenty applications and will then be able to count the "definite words" in a 100-word sample in one to two minutes. The application of the rest of the readability test will ordinarily take another one to two minutes.

Received September 26, 1950.

Early publication.

References

1. Boder, D. P. The adjective-verb quotient: a contribution to the psychology of language. *Psychol. Record*, 1940, 22, 310-343.
2. Curme, George O. *Principles and practice of English grammar*. New York: Barnes & Noble, 1947.
3. Dale, Edgar, and Chall, Jeanne S. A formula for predicting readability. *Educ. Res. Bull.*, Ohio State Univ., 1948, 27, 11-20, 28.
4. Flesch, Rudolf. *The art of readable writing*. New York: Harper & Brothers, 1949.
5. Flesch, Rudolf. A new readability yardstick. *J. appl. Psychol.*, 1948, 32, 221-233.
6. Frank, Jerome. *Courts on trial; myth and reality in American justice*. Princeton Univ. Pr., 1949.
7. Fries, Charles C. One stylistic feature of the 1611 English Bible. *Fred Newton Scott Anniversary papers*. Univ. of Chicago Pr., 1929, 175-187.
8. Hayakawa, S. I. *Language in thought and action*. New York: Harcourt, Brace & Co., 1949.
9. Jespersen, Otto. *Essentials of English grammar*. New York: Henry Holt & Co., 1933.
10. Johnson, Wendell. *People in quandaries; the semantics of personal adjustment*. New York: Harper & Brothers, 1946.
11. LaBrant, Lou L. A study of certain language developments of children in grades 4 to 12 inclusive. *Genet. Psychol. Monogr.*, 1933, 5, 387-491.
12. Lorge, I. The Lorge and Flesch readability formulae: a correction. *Sch. & Soc.*, 1948, 67, 141-142.
13. McCall, W. A., and Crabbs, Lelah M. *Standard test lessons in reading*. Books II, III, IV, and V. New York: Bur. of Publ., Teachers Coll., Columbia Univ., 1926.
14. Perrin, Porter G. *Writer's guide and index to English*. Chicago: Scott, Foresman & Co., 1942.
15. Piaget, Jean. *The child's conception of causality*. New York: Harcourt, Brace & Co., 1930.
16. Piaget, Jean. *Judgment and reasoning in the child*. New York: Harcourt, Brace & Co., 1928.
17. Piaget, Jean. *The language and thought of the child*. New York: Harcourt, Brace & Co., 1926.
18. Séchéhaye, Albert. *Essai sur la structure logique de la phrase*. Paris: Champion, 1926.
19. Shannon, C. E. A mathematical theory of communication. *Bell System Tech. J.*, 1948, 27, 379-423, 623-656.
20. Weaver, Warren. The mathematics of communication. *Scientific American*, July 1949, 181, No. 1, 11-15.
21. Whyte, Lancelot L. *The next development in man*. New York: Henry Holt & Co., 1948.

Personality Characteristics of Embalmer Trainees

Daniel N. Wiener and Werner Simon, M.D.

*Veterans Administration, Fort Snelling, Minnesota**

Occasionally clients are counseled who have a strong desire to be embalmers or undertakers. This vocational objective is seldom suggested to a client unless he shows high original motivation to enter this field.

In a regional veterans' counseling program thirty-six veterans indicated their desire to become embalmers and began vocational training. The vocation is recognized through standards set by the State of Minnesota, and the organized training program includes two years of college work prior to specialized training.

Psychiatrists and psychologists, in discussing the occupation of embalming, frequently raise the question as to personality traits of those attracted to this field. In an attempt to investigate this question a study has been made of the personality characteristics of veterans who indicated their interest in this work, in order to determine whether they are obviously unique in personality. This is but one of many occupations where the technical skill and social attitudes involved may be less important factors to the counselee than his personal peculiarities in motivation and needs.

Literature

In a review of the recent literature of psychiatry and psychology for possible prior research in this area, no study of any kind has been found, although there are a number of analyses on concepts of death, as well as several case reports on necrophilia. The literature on death concepts has only secondary importance here. Perhaps it is sufficient to say that the fear of death, which appears so commonly in reports of dreams and conscious anxieties, is seldom discussed openly. This apparently repressed concern is indirectly reflected, however, in a number of ways; e.g., public responsiveness to "fear" campaigns to raise funds for medical research (cancer, infantile paralysis, tuberculosis), and in the elaborate

rituals of funerals. In relation to the present problem this literature on concepts of death suggests an investigation of the personality structure of those who seem to lack this common fear as shown by their work with the dead (2, 3).

Case reports on necrophilia stress pathological personality organization, especially in the sexual sphere. Brill (1) reports the case of a thirty-year-old single male who was primarily a passive homosexual and whose necrophilia represented a by-product of his pathological sexuality. He gave a history of a severe phobia with regard to corpses since early childhood, which was apparently overcome by touching a corpse. This experience changed the former phobia into the opposite, a strong attraction for corpses. As his fascination for corpses increased, he became the paid assistant of an undertaker, who spoke highly of his diligence and skill. During analysis he confessed that it was his morbid desire to play with corpses which prompted him to work for the undertaker.

Rapoport's (4) patient was a fifty-year-old man who was arrested because he was found kissing and touching a female corpse in a funeral parlor. He began his practice of visiting dead bodies six years after the death of his mother, and a ritual of visiting funeral parlors two or three times a week ensued. The corpses were usually female, middle-aged or elderly. Rapoport points out that the literature contains very little on the subject of necrophilia, but it seems to him unlikely that the number of cases is as small as the absence of literature suggests.

Method

It seemed desirable to determine common impressions of psychologists and psychiatrists as to what personality peculiarities, if any, embalmers might have. This was done by asking both vocational and personal counselors, as well as psychiatrists, for their impressions, and tabulating their responses.

* Published with permission of the Dean's Committee, Medical School, University of Minnesota.

The actual personality characteristics of embalmer trainees were measured by use of the Minnesota Multiphasic Personality Inventory, with the recently developed subtle and obvious keys (5) included. A group of thirty-six veterans were studied, all of whom indicated a desire to be embalmers and made this their training objective after counseling. The control group consisted of one hundred veterans who came to the same guidance centers for counseling during the same period of time, but who decided upon other vocational objectives. All were tested with the MMPI.

The use of the term "embalmer" may be distinguished from that of "undertaker." The undertaker receives training first in embalming, followed by additional training involving more of the business and social details of funeral arrangements.

Results

Thirteen vocational and personal counselors were asked to write briefly their impressions of the personality characteristics of men interested in embalming. Thirty-six responses were elicited, with a range of one to five reactions from each counselor. Table 1 indi-

Table 1

Personality Characteristics of Embalmers According to Thirteen Personal and Vocational Counselors

Traits	Frequency
Psychopathic	11
High Drive	9
Feminine	5
None, Stable	5
Introvertive	3
Scientific Interest	2
Depressive	1
	—
	36

cates the results of the tabulation of the responses. Relatively shallow emotionality, as well as aggressiveness, are the outstanding impressions, with feminine characteristics also suggested.

Five psychiatrists gave ten responses to a similar question, with a range of one to three items. On tabulation, responses classified as "extrovertive," "compulsive," "passive-aggres-

Table 2

Means and Standard Deviations of Embalmer (N = 36) and Control (N = 100) Groups in Age and Education, and on MMPI

		Mean	S.D.
Age	Emb.	25.72	4.66
	Con.	24.48	4.24
Education	Emb.	12.00	1.55
	Con.	11.68	1.63
K	Emb.	56.45	7.85
	Con.	55.15	9.60
F	Emb.	50.20	4.90
	Con.	50.95	5.80
Hypochondriasis*	Emb.	54.25	8.55
	Con.	51.00	9.55
Depression**	Emb.	49.65	10.05
	Con.	52.35	11.40
D-Subtle	Emb.	48.70	11.75
	Con.	51.30	9.20
D-Obvious	Emb.	51.30	9.20
	Con.	55.90	10.60
Hysteria	Emb.	55.90	8.55
	Con.	55.35	9.50
Hy-Subtle	Emb.	57.70	10.50
	Con.	52.05	8.40
Hy-Obvious	Emb.	52.05	11.80
	Con.	55.50	8.95
Psychopathic Dev.	Emb.	55.50	8.15
	Con.	55.54	8.70
Pd-Subtle	Emb.	55.50	7.45
	Con.	56.90	8.50
Pd-Obvious	Emb.	50.35	8.55
	Con.	50.75	10.90
Femininity**	Emb.	47.85	6.15
	Con.	50.75	6.75
Paranoia*	Emb.	47.85	10.10
	Con.	50.75	8.00
Pa-Subtle	Emb.	51.75	9.45
	Con.	44.00	11.90
Pa-Obvious	Emb.	44.00	11.45
	Con.	55.10	11.70
Psychasthenia	Emb.	55.10	9.80
	Con.	54.44	2-3%
Schizophrenia	Emb.	56.35	8.00
	Con.	56.14	9.45
Hypomania**	Emb.	59.55	11.90
	Con.	56.80	11.45
Ma-Subtle	Emb.	61.30	11.70
	Con.	51.75	9.80
Ma-Obvious	Emb.	51.75	9.80
	Con.	51.75	9.80

* Difference between means is significant at 2-3% level.

** Difference between means is significant at 6% level.

sive" and "none" were given twice. There were single responses for "depressive" and "psychopathic."

Table 2 gives the means and standard deviations for the embalmer trainees and control groups, together with an indication of the significance of the difference between the means of each MMPI Scale.

In age and education the embalmer and control groups were not significantly different from each other. In Hypochondriasis, Depression, Femininity, Paranoia and Hypomania, however, the means do differ slightly from each other with some likelihood of significance. Most prominent is the slightly greater tendency toward Hypochondriasis in the embalmers. Hypomania is somewhat less elevated, while in Depression, Femininity and Paranoia the embalmers are slightly lower than the control group. Most of the subtle scores are higher than the obvious scores, a characteristic of successful groups in previous studies (6).

Summary

These test results indicate that there are no strongly deviate personality characteristics among the embalmer trainees. Elevations on the subtle scores, along with the Hypomania score, seem to characterize successful groups in other areas (6). The fact that the scores for

several other scales are somewhat lower for embalmers than for the control group suggests the same conclusion.

The strongest possibility for an interpretation of uniqueness lies in the elevated Hypochondriasis. Tendencies toward over-concern with personal health may, through projection, lead to an interest in the dead and in preparing bodies for burial.

Negatively, the results of this study do not substantiate tendencies toward shallow emotionality, femininity, compulsiveness or necrophilia, which a strong interest in dead bodies might suggest and which both counselors and psychiatrists tend to indicate.

Received January 23, 1950.

References

1. Brill, A. A. Necrophilia. *J. crim. Psychopath.*, 1941, 2, 433-443.
2. Caprio, F. S. A psycho-social study of primitive conceptions of death. *J. crim. Psychopath.*, 1943, 5, 303-317.
3. Orlansky, H. Reactions to the death of Pres. Roosevelt. *J. soc. Psychol.*, 1947, 26, 235-266.
4. Rapoport, J. A case of necrophilia. *J. crim. Psychopath.*, 1942, 4, 277-289.
5. Wiener, D. N. Subtle and obvious keys for the Minnesota Multiphasic Personality Inventory. *J. consult. Psychol.*, 1948, 12, 164-170.
6. Wiener, D. N. A control factor in social adjustment. *J. abnorm. soc. Psychol.* (In press.)

The Guilford-Zimmerman Temperament Survey and Certain Related Personality Tests *

Claudia Gilbert

The Pennsylvania State College

As a result of an accumulation of evidence obtained by testing supervisors in industrial plants in Pennsylvania, the Personnel Service Division of The Pennsylvania State College is now using the Bernreuter Personality Inventory and the Guilford-Martin Personnel Inventory in test batteries for the selection of candidates for supervisory positions. Consequently, the publication of a new personality test, the Guilford-Zimmerman Temperament Survey, which appears to incorporate in one test the traits now being measured by a combination of the Guilford-Martin and the Bernreuter, is of great interest to the Division.

The purpose of the present investigation was to set up an experimental situation which allowed a comparison of the Guilford-Zimmerman, Guilford-Martin, and Bernreuter, and an analysis of the resulting intercorrelations, in order that the Personnel Service Division might have a starting point for further experimentation with this new test.

The Tests

Since the purpose of this study was to make a practical comparison of the Bernreuter with the Guilford-Zimmerman, it was decided to use the three most widely applied Bernreuter scales—B1-N-Neurotic Tendency (reversed and oriented as Stability), B2-S-Self-Sufficiency, and B4-D-Dominance—these also being the scales used by the Personnel Service Division in selection for supervisory trainees.

The Guilford series of personality tests are the result of an extensive series of factor analyses that have identified a number of personality factors. Of the original Guilford series of three personality scales, the Guilford-Martin Personnel Inventory measuring Objec-

tivity, Agreeableness, and Cooperativeness has been found to offer good possibilities in industrial situations. The authors have published studies which give an indication of its validity, particularly in regard to the selection of supervisors (5).

The new Guilford-Zimmerman Temperament Survey measures ten Guilford traits termed GRASE-OFTPM. These ten traits, measured with 300 items by the Guilford-Zimmerman, have been arrived at by means of condensations and omissions of trait scores where intercorrelations are sufficiently high, thus purporting coverage in a more economical fashion of the 13 factors measured by the three original Guilford inventories. Rather than being composed of questions, the new inventory is made up of declarative statements, so that instead of "Do you get things done in a hurry?" it is, "You get things done in a hurry." No validation studies of the Guilford-Zimmerman Temperament Survey are known to the author.

Procedure

The subjects for the study were the members of two introductory psychology classes at The Pennsylvania State College. Of these, an all male population of 64 was used. Scatter diagrams were prepared in order to ascertain the linearity of the data. The Pearson product moment formula was employed and, in all, 120 correlations were determined.

Results

Intercorrelations of the Traits. In Table 1 are presented the coefficients of correlation, in terms of product moment r 's, between the Guilford-Zimmerman factors and their counterparts in the Guilford-Martin. O-Objectivity in the Guilford-Zimmerman correlated .78 with O-Objectivity in the Guilford-Martin; F-Friendliness in the Guilford-Zimmerman correlated .76 with Ag-Agreeableness in the Guilford-Martin, and P-Personal Relations in

* This material is derived from the author's thesis submitted in June, 1950, in partial satisfaction of the requirements for the M.S. degree at The Pennsylvania State College. The author is deeply indebted to Charles Griffin and Dr. Kinsley R. Smith for their guidance and criticism.

Table 1
Coefficients of Correlation *

		Guilford-Zimmerman				
		O-Objectivity	F-Friendliness	P-Personal Relations	E-Emotional Stability	A-Ascendancy
Guilford-Martin	O-Objectivity	.78				
	Ag-Agreableness		.76			
	Co-Cooperativeness			.82		
Bernreuter	B1-N (rev.) Stability				.80	
	B4-D-Dominance					.80

* All r 's significant at the .01 level.

the Guilford-Zimmerman correlated .82 with Co-Cooperativeness in the Guilford-Martin.

Also in Table 1 are presented the coefficients between the Guilford-Zimmerman factors, E-Emotional Stability and A-Ascendancy, and traits B1-N (reversed) Stability and B4-D-Dominance in the Bernreuter. E-Emotional Stability in the Guilford-Zimmerman correlated .80 with B1-N (reversed) Stability in the Bernreuter, and A-Ascendancy in the Guilford-Zimmerman correlated .80 with B4-D-Dominance in the Bernreuter.

Further Analysis of Intercorrelations. In extending the comparison between the factors in the Guilford-Martin and their counterparts in the Guilford-Zimmerman to include comparison with the remaining seven factors in the Guilford-Zimmerman as well as with the Bernreuter traits, several correlations fell into the category of r 's generally accepted as denoting high to very high relationship (4). A correlation of .74 was found between O-Objectivity in the Guilford-Martin and B1-N (reversed) Stability in the Bernreuter. Also, there was a correlation of .74 between this same factor O-Objectivity in the Guilford-Martin and E-Emotional Stability in the Guilford-Zimmerman.

In comparing the Bernreuter traits with the Guilford factors, the only correlation between

these measures which could be accepted as high, aside from those already discussed, was the .74 coefficient between B1-N (reversed) Stability in the Bernreuter and O-Objectivity in the Guilford-Martin.

The intercorrelations found in this study when comparing the Guilford-Zimmerman traits to each other, in terms of product moment r 's, indicated similar interrelationships to those indicated by the tetrachoric coefficients published in the Guilford-Zimmerman manual. There was one very high coefficient, however, which fell well above our arbitrary standard for a high correlation. This was a .84 between E-Emotional Stability and O-Objectivity.

The high intercorrelation between B1-N (reversed) Stability and O-Objectivity in both the Guilford-Martin and the Guilford-Zimmerman brings to mind the high intercorrelation of -.95, as reported in the Bernreuter Manual, between B3-I-Introversion and B1-N-Neurotic Tendency. This suggests, as does the data presented in this paper in terms of correlation coefficients, that on the one hand B1-N, oriented as Stability and B3-I, oriented as Extraversion, measure the same thing, and on the other hand, that this is the same trait as the Guilford-Zimmerman and the Guilford-Martin factor O-Objectivity. The high correlation of .84 between O-Objectivity and E-Emotional Stability in the Guilford-Zimmerman again suggests that these are not separate traits.

To sum up, the evidence presented suggests that the various personality inventories which are measuring traits named Objectivity (or Extraversion) and Emotional Stability, are primarily measuring only one factor.

A Comparison of B2-S-Self-Sufficiency with the Guilford Factors. There was no indication of a relationship between B2-S-Self-Sufficiency and any of the Guilford factors. The highest correlation was, in fact, .37, which though significant at the .01 level of confidence, appeared to be more noteworthy for its lowness than for its highness. However, such a result seems understandable, since up to this point the trait Self-Sufficiency has not been named by Cattell, Guilford, or Thurstone, and apparently has not been isolated in factor analysis. Also, it remains the most indepen-

dent measure, in terms of intercorrelations, in the Bernreuter Personality Inventory.

On the other hand, there could be another explanation. An inspection of the scatter diagrams of B2-S with the Guilford-Zimmerman factors suggests that in regard to several factors, especially A-Ascendancy, there may be curvilinearity rather than linearity of relationship. In terms of the present research, this means the relationship between the Guilford factors and B2-S may have been masked when studied in terms of the Pearson r , and significantly higher correlations might have appeared if the *eta* correlation were applied. To check this assumption the *eta* between B2-S-Self-Sufficiency and A-Ascendancy was computed. The chi-square test for linearity of regression indicated, however, that the departure of curvilinear relationship from linear relationship was not significant at the .05 level.

Summary

A study was made of the relationships between the factors measured by the Guilford-Zimmerman Temperament Survey and the traits measured by two previously validated inventories, the Guilford-Martin Personnel Inventory and the Bernreuter Personality Inventory. The subjects for the study were members of two introductory psychology classes at The Pennsylvania State College. Of these, an all male population of 64 was used. Scatter diagrams were prepared in order to ascertain the linearity of the data. The Pearson product moment formula was employed and, in all, 120 correlations were determined.

1. O-Objectivity in the Guilford-Zimmerman has a high positive correlation ($r = .78$) with O-Objectivity in the Guilford-Martin.

2. P-Personal Relations in the Guilford-Zimmerman has a high positive correlation ($r = .82$) with Co-Cooperativeness in the Guilford-Martin.

3. F-Friendliness in the Guilford-Zimmerman has a high positive correlation ($r = .76$) with Ag-Agreeableness in the Guilford-Martin.

4. E-Emotional Stability in the Guilford-Zimmerman has a high positive correlation ($r = .80$) with B1-N-Neurotic Tendency reversed and oriented as Stability, in the Bernreuter.

5. A-Ascendancy in the Guilford-Zimmerman has a high positive correlation ($r = .80$) with B4-D-Dominance in the Bernreuter.

6. There is no indication of a high degree of relationship between B2-S-Self-Sufficiency on the Bernreuter and any of the Guilford factors.

7. The high positive correlation between the Guilford-Zimmerman factors O-Objectivity and E-Emotional Stability of .84 suggests that these are not separate traits. There is further evidence for this in the high positive correlation between O-Objectivity in the Guilford-Martin and E-Emotional Stability in the Guilford-Zimmerman of .74, and between O-Objectivity in the Guilford-Martin and B1-N-Neurotic Tendency reversed and oriented as Stability in the Bernreuter of .74.

Received June 26, 1950.

Early publication.

References

1. Allport, G. W. Neurotic personality and traits of self-expression. *J. soc. Psychol.*, 1930, 1, 524.
2. Bernreuter, R. G. *Manual for the Personality Inventory*. Stanford University, California: Stanford University Press, 1935.
3. Cattell, R. B. Confirmation and clarification of the primary personality factors. *Psychometrika*, 1947, 12, 197-220.
4. Garrett, H. E. *Statistics in psychology and education*. New York: Longmans, Green and Company, 1948.
5. Guilford, J. P. *Manual for the Guilford-Martin Personnel Inventory*. Beverly Hills, California: Sheridan Supply Company, 1943.
6. Guilford, J. P., and Zimmerman, W. S. *Manual for the Guilford-Zimmerman Temperament Survey*. Beverly Hills, California: Sheridan Supply Company, 1949.
7. Mosier, C. I. A factor analysis of certain neurotic tendencies. *Psychometrika*, 1937, 2, 268-287.
8. Super, D. E. The Bernreuter Personality Inventory: a review of research. *Psychol. Bull.*, 1942, 39, 94-125.
9. Thurstone, L. L. The factorial description of temperament. *Science*, 1950, 111, 454-455.

The Non-Respondent Problem in Questionnaire Research

W. Leslie Barnette, Jr.

Department of Psychology, University of Buffalo

A great deal of social science research is necessarily conducted by means of mail questionnaires, despite wide recognition of pitfalls with such methodology. It is imperative, especially when questionnaire returns obtained on a sample are to be utilized for predictions for a population—the usual goal—that information be available about the non-respondents.

No attempt will here be made to review the literature which discusses this topic. The interested reader is referred to three recent summaries (5, 6, 12). Most workers attempt to minimize non-respondent bias by maximizing their returns, usually through repeated follow-ups. Toops' early reports (9, 10) are well known in this connection. Here he was interested in securing a 100 per cent response, something which may rarely be attempted when the original sample is very large. Others (3, 7, 8) have tried to cope with the problem by using special appeals and even bribes. Despite the fact that these techniques are effective, one is still left with a sizeable "hard core" non-respondent group. It becomes imperative to assess the possible source of bias here.

It is not always the case that the more verbal individuals respond. Clark's study (4), involving a mail questionnaire to A. F. L. union members, demonstrated this. The experience of the present writer has been similar (1), during a pre-testing of a questionnaire to be used on a large group of veterans. Here some 275 questionnaires were mailed out to a random sample; three weeks later a 50 per cent response was secured with no intervening follow-up. Just as many replies were received from veterans in professional and managerial categories as from those in skilled and semi-skilled groups.

In the opinion of the writer, the only *real* solution is by taking a random sample of the entire non-respondent group. By repeated and persistent follow-ups one aims to secure, ideally, a 100 per cent return. One then has relevant data for the detection of non-respondent

bias. The following experience of the writer is offered as evidence.

A large-scale follow-up of some 1,300 veterans in the Greater New York area was in progress. Replies to the first questionnaire mailing totalled 580; a second mailing, some three months later, produced 310 additional replies. The combined first two waves thus gave 890 (69 per cent) respondents. The average time lapse between the date last seen and the questionnaire receipt was two years. All addressees had received free educational and vocational counseling at the agency in question so that they were presumably in its debt. First versus second waves of returns, when checked by chi-square, showed no important differences. However, despite this comforting fact, there still remained a total of 409 (31 per cent) non-respondents.

It was not practical to attempt a third contact of the entire 409. A random sample of the group was drawn (every fifth name from an alphabetical file, which gave $N = 82$). It was not then a matter of merely sending out a third letter and waiting for the subsequent wave of returns. Persistent attempts were necessary with this group by means of intervening postcard reminders, further letters, telephone calls where feasible and, in some cases, personal calls at homes.

Parenthetically, the unit cost of such a procedure is prohibitive. For many of this group, a total of five separate personal letters were involved, each time inclosing the questionnaire and a stamped, self-addressed envelope; for others, suburban New York telephone calls, two and three per person. Such could rarely be done for as large a sample as the entire 409 recalcitrants. The net result was an expenditure of a great amount of time and effort and money for, in the main, a relatively small amount of data in return.

The spontaneous comments of some of these individuals showed that the repeated reminders by mail, the tone of which became more sharp each time, were effective. One respondent,

apparently ashamed by the persistent reminders, finally returned the completed questionnaire, after receipt of the fifth letter, together with five three-cent stamps, as if in payment for the previously unsuccessful efforts at contact.¹ Another respondent, after receipt of the sixth letter, returned the completed questionnaire with an apologetic note: "... thanks for having such patience with me and for making me ashamed of myself to a degree where I find myself resolving not to procrastinate in any form or manner whatsoever." Few, however, were as courteous or gave any signs of such remodelled behavior.

In the end, returns were received from all but three persons in this random sample (i.e., 96 per cent replies). The results of this non-respondent check need not be reported here. Suffice it to say that, in the main, the trends observed in the data from the first two waves were corroborated. A few important differences were evidenced, however, which permitted a more reliable estimate for the entire veteran sample than would otherwise have been possible. These results are presented in two forthcoming papers by the present writer (2).

A subsidiary check for *respondent* bias (i.e., "early" versus "late" returns), which might be a cue for the presence of non-respondent bias, is occasionally available to the researcher. It is never one that might be employed on a large scale, however, without involving considerable annoyance to addressees. When one is engaged in sending out a second and third letter to non-respondents, it sometimes happens that one draws two returns from the same individual. One may then compare these two sets—a type of test-retest technique. The present writer met this situation for only five respondents; there was a minimum of three months' time lapse between the receipt of the first and second questionnaire. The average number of items on the questionnaire responded to by these persons was 12. The mean per cent of exact duplication in answers was 92. The instances of discrepancies between replies from the first versus second questionnaire all related to salary data, all of

which involved a salary raise for the intervening time.

Summary

It is suggested that the real solution to the problem of bias detection, in the event that a numerically large non-respondent group remains after two or three mail questionnaire follow-ups, is to make up a random sample of this hard core; then to make repeated and persistent attempts at contact. The aim here is a 100 per cent response. An illustration of the application of this technique has been given in connection with a large-scale follow-up of 1,300 veterans where, after two mailings, there still remained some 400 non-respondents. An incidental check, of the test-retest type, is pointed out.

Received February 28, 1950.

References

1. Barnette, W. L., Jr. Preliminary report on a follow-up of veterans counseled at the Vocational Service Center, Y. M. C. A. of the City of New York. *Counseling* (Y. M. C. A., 347 Madison Avenue, New York City 17), 1948, 6 (no. 2), 1-2.
2. Barnette, W. L., Jr. Report of a follow-up of counseled veterans. I. Public Law 346 versus Public Law 16 clients. II. Status of pursuit of training. *J. soc. Psychol.* (forthcoming).
3. Bevis, J. C. Economical incentive used for mail questionnaire. *Publ. Opin. Quart.*, 1948, 12, 492-493.
4. Clark, K. E. A vocational interest test at the skilled trades level. *J. appl. Psychol.*, 1949, 33, 291-303.
5. Clausen, J. A., and Ford, R. N. Controlling bias in mail questionnaires. *J. Amer. stat. Ass.*, 1947, 42, 497-511.
6. Norman, R. H. A review of some problems related to the mail questionnaire technique. *Educ. psychol. Measmt.*, 1948, 2, 235-248.
7. Rollins, M. The practical use of repeated questionnaire waves. *J. appl. Psychol.*, 1940, 24, 770-772.
8. Sletto, R. Pretesting of questionnaires. *Amer. sociol. Rev.*, 1940, 5, 193-200.
9. Toops, H. A. Validating the questionnaire method. *J. Person. Res.*, 1923, 2, 152-169.
10. Toops, H. A. The returns from follow-up letters to questionnaires. *J. appl. Psychol.*, 1926, 10, 92-101.
11. Toops, H. A. Questionnaires. Pp. 948-951. In Monroe, W. S. (ed.): *Encyclopedia of Educational Research*. New York: Macmillan, 1950 (revised).

¹ The respondent was in error by 50 per cent; he should have returned thirty cents.

Reactions of Veterans to Counseling *

W. Leslie Barnette, Jr.

Department of Psychology, University of Buffalo

As part of a large-scale follow-up of counseled veterans by means of a mail questionnaire, the reactions of such clients to the counseling process were solicited. These data provide considerable information regarding the value of counseling as seen from the vantage point of the client and have diverse implications for the psychological counselor. At the bottom of a one-page questionnaire, designed primarily to elicit facts about the veteran's current status in training some two years after last seen, two items were placed which directly touched on client attitudes to services received. A pilot study, in which the questionnaire had been pre-tested, produced many letters from veterans amplifying the information requested on the questionnaire itself and dealing specifically with counseling reactions. These proved to be of such value that, when it came to the main study, efforts were made to encourage as many of these spontaneous comments as possible.

Two mailings of the questionnaire were involved. The working total (after deduction of post office rejects) was 1,299. From the two mailings, a total of 890 respondents was secured, containing approximately twice as many non-disabled veterans (those covered by the provisions of Public Law 346) as disabled (Public Law 16). To check on the reliability of the obtained data from these two follow-ups, a random sample of the non-respondents was selected for persistent and repeated follow-up. A 96 per cent return from this non-respondent sample was secured, the results of which altered the main trends of the data in no important way. It is therefore felt that the data from the study of the 890 respondents from the first two waves possess considerable reliability.

The respondents were all completed advise-

ments from the files of the Vocational Service Center of the Y. M. C. A. of New York City (hereafter, VSC). The two questions dealing with attitudes towards counseling were adapted from the reports of the Adjustment Service (6, 10). The first of these asked: Did you consider the suggestions made to you at VSC to be (check one) helpful, of doubtful value, impractical? The second: I found the attitude of the VSC staff to be (check one) sympathetic, indifferent, discourteous. The distribution of the replies from the 890 returns is given in Table 1.

Table 1

Veterans' Opinions of VSC

Suggestions were:		
Helpful	668	(75%)
Doubtful	169	(19%)
Impractical	28	(3%)
No reply	25	(3%)
Totals	890	100
Staff attitude was:		
Sympathetic	809	(91%)
Indifferent	56	(6%)
Discourteous	0	
No reply	25	(3%)
Totals	890	100

Despite the highly favorable opinion, replies falling into the other categories were welcomed as evidence that the returns were not producing a kind of "white-wash." Many respondents appended further comments in support of this majority opinion, two samples of which are:

Glad I visited your office. Put me on the goal to my future. If not for your counselors I guess I never would enjoy working today. Thanks for your kind service.

... Personally I found your counseling so impressive that I resolved to "stick to it" regardless of future distractions, inducements, etc., believing that would be the best, most rewarding, long-range plan.

A total of 19 per cent indicated the counselor suggestions were of doubtful value. These

* This article is based on a section of the author's Ph.D. dissertation in the Department of Psychology, New York University, entitled "Occupational Aptitude Patterns of Counseled Veterans." All statements relating to veterans are the sole responsibility of the author. Colonel Walter Ketcham, Chief of Advice-ment of the New York Regional Office of the Veterans Administration, approved the entire research project.

persons, in the main, either regarded the suggestions as unrealistic because the training suggested would be too extensive or else they indicated disappointment at being given no "new" information about themselves.

A still smaller group (3 per cent) checked the "impractical" category. In support of these ratings one states:

. . . Suggestions impractical because the suggested goal requires years of night study of an extremely complex subject (law) without any possibility of an adequate living for the years of study plus other years of attempting to establish a clientele; and I was and am unwilling to begin life at 40 . . . (continues about his present job). . . . Am doing work I enjoy, am making a living and have time to worry about the world and its works. In other words you can put me down as "adjusted" even though your recommendation was not followed. Thanks for your help.

Turning to a brief consideration of client reactions to the item concerning the staff attitude, the overwhelming majority is again highly favorable. Many of the unsolicited comments were overly enthusiastic, containing direct compliments and testimonials to counselor effectiveness. The following quotations, from the many, will give the flavor.

I thank you and the staff from the bottom of my heart. May there be more people like you. Keep up the good work.

Was most favorably impressed by the helpful attitude and results. At no time was the impression given that here was a group of magicians who could determine an exact endeavor which could not fail—yet the pattern made by the apparently exhaustive tests was very clear. Can't suggest anything more helpful than the hope you'll continue—as is.

A very small percentage indicated they felt some indifference. This is rarely explained and, when attempted, is somewhat garbled. No respondent checked the "discourteous" entry—possibly too strong a word here.

More productive from the point of view of studying counseling from the vantage point of the client are the spontaneous remarks in connection with these two items.¹ This type of

¹ The covering letter which accompanied the questionnaire specifically requested any type of additional information, such as feelings about VSC, that the veteran would like to supply. The letter also encouraged frankness and used the altruistic appeal that such information would be of great value in helping VSC deal with current veteran clients.

supplementary data, of course, was not received from all. Of the total of 890 returned questionnaires, 575 individuals or 64 per cent offered additional remarks. These comments run a large gamut. They vary in length from a succinct thanks for help to a 13-page letter, from notes full of errors in grammar and spelling to typed biographical essays, from a warm and receptive attitude to the questionnaire and the whole idea of a follow-up to irate negativistic attitudes,² from well-adjusted and secure individuals to persons with signs of aggravated emotional instability. All these comments were not limited to counseling matters; it will

Table 2
Content Analysis of All Spontaneous Remarks
(N = 575)

Info. augmenting Q items (i.e., no new facts)	392 (70%)
Incr. self-confidence	80 (14%)
Specific VSC compliments	100 (17%)
Thanks for your interest	152 (26%)
VSC criticisms	102 (18%)
VSC recommendations made	25 (4%)
Preoccupied with tests	89 (15%)
School difficulties	64 (11%)
Job difficulties	65 (11%)
Employer gripe	16 (3%)
Feels goal unrealistic	29 (4%)
Raises questions for answer	64 (11%)
VA criticisms	12 (2%)
Miscellaneous	69 (12%)

be only these types, however, that are discussed in this paper. Most of the commenting group supplied amplification concerning their job or educational history; others gave extensive occupational information; others raised questions that demanded answers; still others simply talked about irrelevancies. Any such classification of these diverse comments must be a subjective one. The categories which proved useful are reported in Table 2. Any one respondent may have made more than one

² The most extreme example of this is from a Public Law 16 client with a 50 per cent disability for psychosis, anxiety. He returned a blank questionnaire with the following comments: "What are you guys doing over there? Pimping for the VA? Last time I had anything to do with you, all I got was a lot of red tape and a 20 per cent cut in my pension. So from now on my name is KIZ MIAZ."

type of comment; the percentages will not therefore total 100.

Increased Self-Confidence

Beyond a general statement from many respondents indicating an appreciation for the interest and help extended, many mention the fact that tests and counseling served to increase self-confidence. To vocational counselors, this is no great surprise. Seipp's report (10) for the Adjustment Service noted the same result; Stone and Simos (11), in their Minnesota follow-up, likewise. Despite the fact that no specific question was asked regarding this point, 80 (14 per cent) of individuals making spontaneous comments specifically mention this point.

Tests encouraged me to continue with pre-war program; confidence increased. . . . Saved the trouble of finding out the hard way.

. . . In helping me realize what action to take, you and your staff opened a bright road in front of me, for the road I had traversed before was always dark and had no hope to see a light at the end of it.

Critical Comments on Counseling

Respondents were not uniformly so commendatory. On the questionnaire some 22 per cent indicated they regarded the suggestions given as of doubtful value or impractical. Some of these went on to a fuller explanation of such attitudes. While many such comments evidence great misapprehension as to the functions of a counseling agency, at least most are honestly offered and, in the main, well meant. These remarks frequently betray an exaggerated feeling of dependence towards psychological help. These respondents feel VSC is not definite enough in its suggestions, while others think it is too definite; others think too much emphasis is placed on interests, claimed or measured. And others are annoyed that the tests and counseling did not provide them with startlingly novel information about themselves. Some feel not enough pressure is brought by the counselor—to "force decisions," as one respondent phrased the matter.

. . . Tests and interviews left me just about where they found me. Confirmed obvious attitudes; showed up my inaptitudes. Got no actual guidance. When left the building, was just as befuddled as to the next step as when

started. No schooling or type of job was recommended. . . . Perhaps in helping a party such as me, it is harder to point out the road to follow. . . . In my case perhaps I had to find the machinery or equipment I *wanted* to sell. . . . Only an experienced person in touch with that type of business could have helped me. . . .

. . . Staff courteous but not convincing enough. Should actually force the individual to try to decide whether he likes an occupation or not. . . . Why not arrange for actually visiting different occupations by potential employees? . . . What good is testing a man whose test reveals he is best suited for a closed or already over-crowded field: spread out—don't limit yourself to one small part of the job.

Recommendations Offered

A few respondents—25 in all or some 4 per cent of the comment group—tried to make specific suggestions as to how the service might be improved. Some of these suggestions come from people who had distinctly favorable attitudes toward VSC but who felt the job might be done still better; other suggestions came from respondents with definitely unfavorable points of view. Actually there were only a small number of constructive suggestions.

The two most frequent recommendations are that the counselors be more specific, do more "telling" and that a placement service be installed. The former is in disagreement with much of the best counseling theory; the latter is now an accomplished fact at VSC. Other suggestions are occasionally sound but there are others that are very unrealistic—those which advocate even more authoritarian approaches to clients. In the former category are suggestions to the effect that more time should be devoted to actual counseling—more discussion of the individual, more "analysis" done. These are the persons who, while they regard the tests as valuable, want to go considerably beyond them. Pressure of time and the case load per counselor frequently operate against this. Certainly, however, wherever this aspect could be lengthened it would be desirable; it would also serve to reduce the ever-present potential of counseling failures.

Test Preoccupied Clients

There has been previous mention of signs of psychological dependency in some respondents,

especially those with negatively critical comments about services rendered. There is another—and much larger—group with some of these same characteristics. These are the clients here designated as “test preoccupied.” In the spontaneous comments offered, some 15 per cent of the group indicated, in one way or another, that the tests were the most valuable part of the service. That is to say, such people mention only this phase of VSC’s work, either favorably or not; they never talk about benefits that might have accrued from counseling. It would appear that the sole benefit—if, indeed, the respondents would term it that—received from VSC was the time spent in the testing rooms. Not all of these are the “low achievers” as described in Friend and Haggard’s monograph (5), but some of them are: people with poor emotional attitudes towards work, individuals with considerable rigidity as regards personality structure. Some, however, have already shown signs of good achievement. But the sheer frequency with which such comments crop up suggests client orientation toward vocational counseling that needs changing.

As Covner (3) and others have pointed out, the usefulness of aptitude tests where it is demanded that counselors provide specific occupational information has long been apparent. An approach that is not exclusively oriented to aptitude testing may easily be applied both to the preparation of the client for tests and to interpretation of such results. The comments of this group evidence the fact that many leave the entire counseling set-up with attitudes about tests that need redirecting.

Others blame the tests for lack of specificity. The tests only told them what they already knew; there were no tests available for some special occupational area, etc. A few respondents specifically refer to the tests, especially the interest inventories, as being too susceptible to “faking.” Several respondents, as if not satisfied with the tests administered, want more. Possibly they have changed their occupational goal but still find themselves undecided and uncertain. In any event, they look to “the tests” for answers. The psychological insecurity of some of these clients is

apparent from their comments. A more exaggerated picture of this is to be seen in clients who have “shopped around” for other tests. The most extreme case is the one respondent whose long letter (13 pages) indicated he had been to a minimum of seven different counseling agencies and had been given tests at all.

It should be borne in mind by the reader that such “test preoccupied” clients only represent a small proportion of the total group (15 per cent). A direct question about the test attitudes of respondents was not asked. Had this been so, a larger number of such replies might well have been received. Stone and Simos (11), for example, did ask such a question and found some 60 per cent with “test oriented” attitudes. Covner (3) found this a prevalent tendency in his clients; Kohn (7) at the University of Washington reports more than 60 per cent of veterans registering at the university’s counseling center said they desired vocational tests.

At any rate, the counseling process has been something less than a success with many of this group who often need a discussion about unconscious emotional attitudes. Others seem to have only a mild interest in improving present circumstances. Small wonder these continue to look to “the tests” for the answers and continue to insist on greater and greater specificity both in test results and counselor suggestions. Any sound counseling approach which recognized motivation as the primary datum would be of great help here—i.e., one which would be organized around the discovery and correlation of the dynamics of the individual with those of the occupational objective that is to be approved.³

Questions Raised

From many veterans the receipt of the questionnaire brought an immediate response and one that entailed considerable correspondence. It was as if the receipt of the questionnaire revived many unanswered questions in this group and, having had one helpful experience, they turned again to the agency. Numerically

³ See Part IV: “A Study of Job Values” in Friend and Haggard (5) for a conceptual basis upon which such a program might be built. McGregor (8) has discussed the problems of industrial leadership and supervision from the same point of view.

the group is not a large one, in the light of all returns, but the extra time involved with these cases was extensive. A total of 64 individuals (11 per cent of the comment group) raised questions demanding some sort of reply. Many desired a written summary of test results, which was duly sent out; many more betrayed feelings of insecurity about present status, some requesting another interview while others left it up to VSC. The breakdown is as follows: 20 (31 per cent) asked for test scores; 31 (48 per cent) indicated further counseling was needed; 8 (12 per cent) placement help was requested; 5 (9 per cent) miscellaneous.

By and large these individuals are the "test preoccupied" people. They are compulsive about the tests they have already taken; they want more in order to bolster feelings of self-confidence. They ask unrealistic questions about schooling or they want extra help in other ways. Some did not even wait to write a letter but came right in for a visit after receipt of the questionnaire.

Implications for the Counselor

What are the implications for the counselor? The comments show that clients have done considerable after-visit evaluation. From this welter of information some specific suggestions for counselors may be garnered. The respondents have provided a picture of the counseling process as seen from their vantage point. It is evident that the typical client arrives at this situation with two dominant sets, namely, that the counselor will tell him what is best and that the tests will provide most of the answers.

Few good counselors, of course, really play an authoritarian role. It would certainly be wise, with the more dependent clients at least, specifically and explicitly to delineate the counselor-client relationship as a cooperative one. The essential purpose of all sessions is clarification and not to inform the client that he is best equipped to pursue, say, bookkeeping versus accounting. It might be well to state at the outset that the arrival at *any* specific answer regarding future occupational goals is unlikely.

The notion that vocational and educational

problems can be answered by tests must be altered. Here might well be applied a tactic so that the counselor secures a picture of what the client feels on this score at the beginning. There are, after all, vocational problems where it would be unnecessary to administer any tests at all. Interview manuals often refer to the "preparation of the client for tests." It is precisely at this stage where the counselor must so structure his remarks that false hopes and illusions about test results are not established in the client. Pursuing this tactic further, the client may frequently be allowed to participate in the actual business of deciding whether certain tests are suitable for his particular case; thus all responsibility is not placed on the counselor. Bordin and Bixler (2) provide excellent suggestions here; Seeman's study (9) on such a self-selection-test technique is also pertinent.

The interpretation of test results to the client frequently demands all the subtlety, tact and skill that a counselor may muster. It is suggested here that it might be a useful device to secure first from the client a "free" response as to how he thought he performed. By such a means the counselor might quickly tell whether he need proceed cautiously and with understatement or otherwise. Often such a "free" response will disclose that the client has already done some accurate self-evaluation which might now be woven into the counseling fabric. Covner (3) has dealt with this aspect; Bixler and Bixler (1) offer detailed quotations and suggestions on this interpretation problem. At this stage, too, the counselor might attempt to forestall unfavorable reactions (frequently met) with clients actually launched on a job or school career or who are going through such services only because of necessary VA approval. The type of interpretation of test data for this sort of client should be quite different from that given to an individual who has come for counseling because of real need.

Many clients are in need of a frank statement about the nature and purpose of interest inventories. The impression is to be avoided that interests, either claimed or measured, are of primary significance. This is frequently just the sort of notion many clients get, possibly because it is relatively easy for the

counselor-client relationship to get under way by first discussing likes and dislikes. At all costs, the final impression, prevalent among the critical respondents, that all the counselor did was to agree with originally stated likes of the client, is to be avoided. Similarly, too complete reliance on high points of interest profiles may be of doubtful validity, as Diamond (4) has recently discussed. In addition, a succinct caution is sometimes necessary about the function of interest inventories and about the possible operation of unconscious sets on the part of clients that might distort scores.

Counselors should take pains to explain clearly the type of norm data used on which raw scores are evaluated. Many veterans of beyond high school level are deflated when they learn they receive above average scores on tests where twelfth grade norm groups are the standard. This is part and parcel of the need of developing local norms.

Several respondents suggested that some sort of a written summary for each client be provided. The idea, on the surface, appears to be a good one. This might easily be a one-page sheet where test scores are indicated in general terms along with a brief description of the test itself. It is suggested that any precise indication of test scores—such as percentile ratings, sigma scores, etc.—be avoided. To the psychometrically unsophisticated person such devices tend to indicate greater accuracy than they possess.⁴

There must be an honest facing on the part of the counselor towards all client handicaps—such as a client's lack of a high school diploma. No sugarcoating, as clients are quick to perceive this type of "dishonesty"; they specifically mention they resent being sent away happy. Here also one might employ a non-directive question or two, so as to get the client to express in his own words what he feels his present handicaps to be. This would, in turn, help to minimize the "telling" on the part of the counselor.

⁴ Such a practice contains hazards; it is not VSC policy. Here clients who desire further interpretation of test results come in for an additional appointment and the test results are then presented verbally. Such a procedure obviously avoids the pitfalls inherent in any conclusions.

As much advance planning should go into the summary interview as possible. Clients show by their comments that they arrive at this stage somewhat "keyed-up." Presentation of facts about test performance is important but is not the main function of counseling. Here is where skillful weaving of test result data into the counseling situation is paramount. If the client has been well prepared for the tests in preliminary interviews, he will not expect the counselor merely to sit back and "tell." This interview might well be the longest as regards actual time. Clients expect, often unwisely, great things here; they are disillusioned by speed, brevity or anything that might be interpreted as vagueness.

A closing suggestion from the counselor that he would be glad to talk again with the client at some future date would be very acceptable to many. To be sure, this is generally assumed—but often only by the agency personnel and is rarely stated explicitly. This then puts the responsibility for a follow-up squarely on the client. The spontaneous comments mention that clients are very receptive to even a mailed questionnaire one or more years later; even this check prompted several to call at the office for interviews.

It must be recognized at the outset that not all comers are fit for counseling, since there may be dominant unconscious emotional forces operating which vitiate all suggestions no matter how carefully thought out and handled by the counselor. A non-directive beginning with clients might well unearth some of these "unfavorables" near the starting line. The more that is known of these dynamics, the better will all vocational and educational counseling become.⁵

Summary

The reactions to counseling of 890 veterans were studied by means of replies to items on a mail questionnaire together with spontaneous comments. The counselor suggestions were found to be genuinely helpful by 75 per cent of the group; 91 per cent reported the staff

⁵ Friend and Haggard have not only stated the problem explicitly but have made a significant beginning towards this goal; see especially pages 110-111 of their monograph (5).

attitude a sympathetic one. Counseling served to increase the self-confidence of many. The prevalence of a "test preoccupied" group is noted. Implications for the psychological counselor are discussed in terms of the comments offered by these veteran clients.

Received February 16, 1950.

References

1. Bixler, R. H., and Bixler, V. H. Test interpretation in vocational counseling. *Educ. psychol. Measmt.*, 1946, 6, 145-156.
2. Bordin, E. S., and Bixler, R. H. Test selection: a process in counseling. *Educ. psychol. Measmt.*, 1946, 6, 361-374.
3. Covner, B. J. Nondirective interviewing techniques in vocational counseling. *J. consult. Psychol.*, 1947, 11, 70-73.
4. Diamond, S. The interpretation of interest profiles. *J. appl. Psychol.*, 1948, 32, 512-520.
5. Friend, J. G., and Haggard, E. A. Work adjustment in relation to family background. *Appl. Psychol. Monogr.*, 1948, No. 16. Pp. 150.
6. Hawkins, L. S., and Fialkin, H. N. *Clients' opinions of the Adjustment Service*. New York: Amer. Assoc. Adult Educ., Adjustment Service Report XII, 1935. Pp. 95.
7. Kohn, N., Jr. Trends and development of the vocational and other interests of veterans at Washington University. *Educ. psychol. Measmt.*, 1947, 7, 631-637.
8. McGregor, C. Conditions of effective leadership in the industrial organization. *J. consult. Psychol.*, 1944, 8, 55-63.
9. Seeman, J. A study of client self-selection of tests in vocational counseling. *Educ. psychol. Measmt.*, 1948, 8, 327-346.
10. Seipp, E. *A study of 100 clients of the Adjustment Service*. New York: Amer. Assoc. Adult Educ., Adjustment Service Report XI, 1935. Pp. 31.
11. Stone, C. H., and Simos, I. A follow-up study of personal counseling versus counseling by letter. *J. appl. Psychol.*, 1948, 32, 408-414.

The Wechsler-Bellevue Intelligence Scale and High School Achievement*

Arden N. Frandsen

Utah State Agricultural College

The validity of the Wechsler-Bellevue Intelligence Scale as a measure of intelligence as indicated by its correlations with other already established tests of intelligence has been reported by several writers. Its correlation with various other tests have ranged from .39 to .93 (25, p. 134). The median of six different correlations with the Revised Stanford-Binet is .87 (8, 16, 22, 25). And Altus and Mahler found a bi-serial r of .52 between W-B IQ's and graduation versus being discharged as inept from the 9th Service Command Special Training Center (1). Tests, however, have multiple validities, which may vary with differences in the particular criteria one may wish to predict. High school counselors may be especially interested in the validity of the W-B, both as a whole and when abbreviated by division into selected subtest groups, for predicting achievement in high school curricula. The data to be presented in this paper should contribute specifically to this question, and by implication, also to an evaluation of the test for college guidance.

Two previously reported investigations have contributed data on the validity of the W-B in predicting academic achievement at the college level, a situation closely related, of course, to the question of the test's validity for predicting high school academic achievement. Anderson (2), for 112 college women whose mean W-B IQ was 118.5, found correlations with average freshman grades of .45 for full scale IQ, .52 for verbal IQ, and of .23 for performance IQ. Sartain (22), for 50 college students whose mean IQ was 117.5, found the corresponding correlations of freshman grades with W-B IQ's to be .53, .58, and .35, respectively. These coefficients compare very favorably, especially for the W-B verbal scale, with the modal range of correlations (.40 to .60) which have been reported between high school

grades and group verbal intelligence tests (6, p. 846).

Subjects. The subjects used in the studies conducted by Anderson and by Sartain were not selected to represent adequately the range of abilities of college freshmen, nor was the area in which achievement was appraised homogeneous. The subjects of the present study comprise practically an entire class (83 out of 90) of relatively superior high school seniors from a mid-west university city whose employed population consists predominantly of university staff and otherwise of people employed in professional and business occupations. A large proportion (83%) of the high school seniors anticipated attending college and were highly motivated to meet entrance requirements. There were variations in the courses pursued, but nearly all students included a basic core of college-preparatory subjects in their courses of study. And because these high school students were living with their own families, within a fairly narrow range of socio-economic status, their general conditions of living were also probably more similar than they were for the college freshmen. The ages (obtained when the W-B was given) of these 83 high school seniors ranged 16 to 19 with a mode of 17. The range in W-B IQ was from 86 to 145, and the mean was 119.8.

Hypotheses. In this study, the data on intelligence and academic achievement will be analyzed to test three hypotheses. 1. With conditions more favorable for predicting achievement (a more complete range of abilities represented, high motivation, more homogeneous curricula, and probably more uniform general conditions of living), validity coefficients even higher than those found by Anderson and by Sartain for college students may be obtained. 2. Because certain subtests may dilute the predictive efficiency of the W-B for a particular purpose, it will be possible, by omitting the subtests showing lowest correlations with our criterion, to find a combination of subtests that will predict this criterion more effectively than the present full scale. 3. As a corollary to

* To Mr. Carl Hammer, principal, and to Mrs. Gladys Issacs, Counselor, both of the West Lafayette, Indiana High School, the writer is very grateful for their interest and generous help in arranging for testing their students and for supplying the achievement data.

hypothesis 2, for testing situations where time economy is important, valid abbreviated W-B scales may be selected wisely according to their separate validities in predicting a particular criterion. Even with drastic deletions from the full scale, the greater saturation of the abbreviation with valid items (for a particular criterion) may compensate for the reduced reliability of the shorter tests.

The first hypothesis is similar to that proposed by Burt and Arps to explain their higher correlation between Alpha intelligence and academic achievement in military academies compared to public high schools, r 's being .39 and .19, respectively. In the academies, they observed, the "general supervision, supervised study, and a system of reward and punishment are more apt to hold the student to his maximum intellectual ability" (3, p. 393). The latter two hypotheses are rendered plausible by the facts that the 11 different subtests of the W-B were selected to test a variety of mental functions and that the 45 intercorrelations reported by Wechsler (25, p. 223) for 10 of the subtests indicate only a moderate degree of communality among the mental functions being measured by the scale. They range, except for only 11 coefficients outside this range, from .3 to .5, with the mode being .4. It is quite probable, therefore, that the separate tests are measuring to some extent different quasi-general aptitudes and/or specific abilities. And the comparability of the 11 W-B subtest scales makes it feasible and convenient to take advantage of this flexible feature of the W-B test.

Procedure. The data of this study include, for 83 high school seniors: (1) average grade-point-ratios earned in an academic curriculum and cumulated from grade 9 through 12; (2) Henmon-Nelson, Form A IQ's; and (3) Wechsler-Bellevue IQ's and separate subtest weighted scores. The W-B tests were administered to the high school students, at different periods during their senior year,¹ by the writer's Purdue University students in In-

dividual Intelligence Testing, after they had demonstrated to the writer adequate skill in administration of the test. Product-moment correlations were determined between grade-point-ratios and the IQ's for both tests and with the W-B subtests, singly and in various combinations.

Results

1. *IQ's and Grade-Point-Ratios.* The data for interpreting IQ's and for evaluating their validity in predicting academic achievement in this high school are presented in Table 1.

Table 1

W-B IQ and Henmon-Nelson IQ Data for 83 High School Seniors and the Correlations of the Tests with Total (3 years) Grade-Point-Ratios

Tests	Means	S.D.	r with G.P.R.
Wechsler-Bellevue, F-IQ	119.8	11.86	.685
Wechsler-Bellevue, V-IQ	115.7	12.96	.69
Wechsler-Bellevue, P-IQ	119.6	10.08	.48
Henmon-Nelson, Form A IQ	112.1	12.84	.52

Taking as a basis of comparison the Henmon-Nelson IQ and grade-point-ratio correlation of .52, which is as high as has usually been reported in the literature (6, p. 845) for similar prognostic relationships, the correlations between W-B IQ's and grade-point-ratios are surprisingly high. As is shown in Table 1, the W-B verbal IQ alone and the full scale IQ are equally good predictors of high school academic achievement, the correlations being .69 and .685, respectively. For the performance scale, the correlation of .48 with grade-point-ratio is much lower, but it is only slightly lower than the corresponding correlation of .52 with the Henmon-Nelson group test IQ's. The relatively high validity of the W-B, especially the verbal scale, in predicting high school achievement, although perhaps surprising in comparison with the results previously reported both for group tests and for the W-B on the college level, is in line with our hypothesis that when the factors of motivation, and uniformity of curriculum and of living conditions are all favorable, higher validities are to be expected.

¹ Ideally in a predictive study the prognostic tests should have been administered at the beginning of the period covered; much data on IQ constancy, however, perhaps justify the assumption that IQ's obtained in the senior year would not have differed significantly from IQ's that should have been obtained two years earlier.

Referring again to Table 1, it should also be noted that the mean IQ of our group is relatively high, being 119.8 for the W-B full scale. Because of this fact and in the light of Goldfarb's contention that the W-B is relatively ineffective in discriminating among a group of superior adolescents (8), it might have been inferred that the test would have proved to be an inefficient guidance device for our population. The correlation results, however, are inconsistent with this assumption; and three other indices indicate adequate discrimination among our population of superior high school students: (1) the IQ's vary in range from 86 to 145, (2) the standard deviation of 11.86 is only 2.64 points narrower than the standard deviation of 14.50 for Wechsler's standardization group of comparable age (25, p. 122), and (3) the distribution of IQ's appears approximately normal, being not markedly bunched toward the highest IQ's. In interpreting IQ's for these superior students, however, local in addition to "national" norms are helpful. For example, an IQ of 120 which corresponds to a Wechsler percentile of 93 is the equivalent of a percentile score of only 52 according to the local norms. In relating IQ's to probable potential achievement within the high school, the local norms are more consistent; but for vocational guidance, where the scope of educational and occupational competition is national, percentiles based on Wechsler's more representative standardization group are, of course, more appropriate.

In passing, it may also be interesting to note the intercorrelations between the Henmon-Nelson and the Wechsler-Bellevue IQ's. That the W-B verbal scale and the H-N measure quite similar traits is indicated by the correlation between them of .80. Between the W-B full scale IQ's and the H-N IQ's the correlation is .70, indicating fair agreement. But the correlation between H-N IQ's and W-B performance IQ's of only .44 indicates that the latter scale is measuring traits somewhat independent from the verbal abilities measured by both the H-N and the W-B verbal tests.

2. *Correlations between Each W-B Subtest and Grade-Point-Ratio.* As a preliminary both to determining the combination of subtests having a maximum correlation with academic achieve-

ment and to selecting promising subtest combinations for abbreviated scales, the mean subtest weights and their correlations with average grade-point-ratios were determined for each of the 10 subtests administered to our population of 80 high school seniors.² These data are presented in Table 2.

In Table 2, attention is first directed to the variations in subtest means. Two notably low scores are Digit Span and Arithmetic, tests which have in common the fact that they both involve numbers and require for successful performance high degrees of attention (espe-

Table 2
Means and Correlations Between Separate W-B Subtest
Weights and Academic Achievement (3 Years
Cumulative Grade-Point-Ratios) for
80 High School Seniors

	Mean* Wt.	r
Verbal Tests		
Information	11.9	.56
Comprehension	13.0	.51
Digit Span	10.3	.45
Arithmetic	10.6	.55
Similarities	12.5	.54
Performance Tests		
P. Arrangement	12.5	.15
P. Completion	12.7	.26
Block Design	14.2	.44
Object Assembly	13.3	.11
Digit Symbol	12.6	.36

* In computing means, lower and upper limits of score of 10, say, were assumed at 9.5 and 10.5, respectively.

cially digits) and concentration. Findings such as those of Lewinski (17) together with the fairly frequent references in the test protocols to comments by the students on feelings of inadequacy on the "number tests" suggest that in several instances efficiency on these two tests may have been impaired for these highly motivated students by temporary conditioned emotional "panics." On the Block Designs test, which involves mainly analytic and synthetic abilities in dealing with abstract elements in multiple form-color relationships,

² The number of subjects is reduced from 83 to 80 because for three students complete separate subtest scores were unavailable.

these superior adolescents achieved their best performance.

More important as a guide in selecting valid subtest combinations, however, are the correlations of the separate subtests with the achievement criterion. On the whole, as might have been expected from coefficients reported in Table 1, the verbal tests are distinctly more valid for this purpose than are the performance tests, with the exception of the Block Designs test. It may be interesting to note that among the relatively high academic achievers, none earned a low Block Design score. However, several of the low achievers earned inconsistently high Block Design scores. It would seem that the kind of "reasoning" measured by this test may be necessary for high academic achievement, but it alone does not guarantee it. For three subtests (Picture Arrangement, Picture Completion, and Object Assembly) the correlations with grade-point-ratios fail to meet the statistical criterion for significance at the 1 per cent level of confidence. The scatter diagram for PC, on which test the ceiling limits weighted scores to 15, indicates that for these subjects this subtest is not discriminative; scores regardless of academic achievement tended to approach the ceiling. For both PA and OA, although the scores were well distributed over the full range, they apparently corresponded with achievement at no part of the range.

3. *Most Valid W-B Subtest Combination.* According to our second hypothesis, W-B subtests which are separately invalid for predicting the criterion may dilute the predictive efficiency of the other more valid subtests in the full scale. Table 2 clearly indicates three such relatively invalid subtests, the three—PA, PC, and OA—have already been mentioned as having statistically insignificant correlations with grade-point-ratios. Moreover, although all of the W-B subtests correlate significantly (.41 to .73) with the full scale, it happens that, except for Digit Span, these three tests also have the lowest correlations (.41 to .61) with the scale as a whole (25, p. 224). Therefore, when these tests are deleted, the remaining tests should have a greater saturation both with whatever is required to predict our criterion (academic achievement) and with the traits which the

scale as a whole are measuring. The combination 10/7 (I + C + D + A + B + Dsy), which includes all of the subtests except PA, PC, and OA, was found to correlate .765 with grade-point-ratios, and thus supports our hypothesis quite well. It is recognized, of course, that we have not demonstrated statistically that this combination is the combination of maximum validity. Table 3, however, shows that compared to either the full scale or to any abbreviated combination which was evaluated, this combination predicts most efficiently our criterion.

4. *Validity of Abbreviated Scales.* To meet the need for an individual intelligence examination in situations where economy of time is important, several abbreviations of the W-B have been proposed and evaluated (4, 7, 9, 10, 11, 12, 13, 14, 15, 18, 19, 21, 23, 24). In these explorations to find satisfactory abbreviations, the usual requirements in addition to economy of time have been: (1) mean of abbreviation approximating closely the full scale mean; (2) abbreviation IQ's correlating well with full scale IQ's; and (3) combinations of subtests likely to yield data which may be significant in the clinical diagnosis of personality. Criterion 1 has usually been satisfactorily attained by simple pro-rating of test weights or as in two instances (4, 15) by special factor weightings. Correlations ranging from .80 to .97 indicate that criterion 2 has been acceptably met. Except in one investigation (11), however, wherein abbreviation "CVS" subtest pattern was found to distinguish statistically both aged and schizophrenic subjects from normal and feeble-minded subjects, the claims that the abbreviations would yield significant "clinical" data have been supported only on the basis of logic and clinical experience. In this study of W-B abbreviations we shall be concerned with another criterion, namely their validity in predicting academic achievement in high school. In selecting promising subtest combinations, we shall, therefore, besides considering the abbreviations already proposed³ and the "aptitude" groupings suggested by

³ Because we did not regularly administer the Vocabulary subtest, some abbreviations (10, 11, 12, 18, 19) which included this excellent test or an abbreviated substitute for it (14, 24) were omitted in our evaluations.

Table 3

Mean IQ's and Correlations with Academic Achievement of the W-B Full Scale IQ, the Most Valid W-B Combination, and 12 Selected W-B Abbreviations for 80 High School Seniors

Combination of Subtests	IQ for Mean Pro-Rated Wts.	<i>r</i> with G.P.R.
a. W-B full scale IQ	119.8	.685
b. 10/7 (I+C+D+A+S+B+Dsy), Highest validity	118.0	.765
1. Geil's 5/2 (C+S+D+B)	121.7	.71
2. Patterson's "Reasoning Group" 5/2 (A+S+B+C)	123	.69
3. Rabin's 10/3 (C+A+S)	118	.69
4. 10/3 (I+A+S), Highest 3 <i>r</i> 's with G.P.R.	114	.69
5. Diamond's "Linguistic Factor" 10/3 (I+C+S)	119	.65
6. Diamond's "Clerical Factor" 10/3 (D+A+Dsy)	118.5	.52
7. Diamond's "Spatial Factor" 10/3 (PC+OA+B)	112	.37
8. 10/3 (C+A+B), Highest <i>r</i> 's plus qualitative clues	121	.62
9. Cummings, MacPhee, & Wright's 5 (C+A)	116	.625
10. Gurvitz's 5 (PA+D)	112	.39
11. 5/2 (I+A), Highest 2 <i>r</i> 's with G.P.R.	111	.63
12. 5 (A+B), Reasoning: Numerical and spatial symbols	120	.60

the factor research of Diamond (5) and Patterson (20), be guided by Table 2 showing the correlations of separate subtests with our criterion. All of these combinations and the data for evaluating them are presented in Table 3.

The data in Table 3 support quite well our third hypothesis, that concerning the possibility of finding W-B abbreviations which are valid for a particular purpose. According to the criterion of validity in predicting academic achievement, the table shows that four abbreviations—numbers 1, 2, 3, and 4—equal the efficiency in this respect of the full scale, although they require only about half the time for administration. Five other three-test or even two-test abbreviations—numbers 5, 8, 9, 11 and 12—besides being even more economical of time, are reasonably efficient in predicting academic achievement, the correlations ranging from .60 to .65. All are superior in this respect to the Henmon-Nelson group test of intelligence. That the abbreviations are not all effective, however, is clearly indicated by the low correlations of .37 and .39 for Diamond's "Spatial Factor" and Gurvitz's "ten-minute" scale. Nevertheless, three two-test abbreviations—numbers 9, 11, and 12—predict academic achievement for these students fairly efficiently, *r*'s ranging from .60 to .63.

Summary

Using as subjects practically an entire class of 83 relatively superior high school seniors, the validity of the Wechsler-Bellevue Intelligence Scale, both as a whole and in various subtest combinations, has been evaluated for effectiveness in predicting academic achievement.

1. Both W-B full scale IQ's and verbal scale IQ's are found to predict three year average grade-point-ratios very efficiently, *r* being .69 for both tests. The W-B performance scale, however, measures traits less significantly related to academic achievement.

2. When the subtests found separately least valid for predicting achievement are deleted from the full scale, the resulting combination, 10/7 (I + C + D + A + S + B + Dsy), is found to correlate even higher (.765) with grade-point-ratios. This is probably the W-B combination of maximum validity in predicting high school academic achievement.

3. By choosing subtest combinations according to their correlations separately with the criterion, at least nine different W-B abbreviations of satisfactory validity are found, correlations with academic achievement ranging from .60 to .71.

Our data are limited to high school seniors,

but in the light of the results obtained by Anderson (2) and by Sartain (22), it is probable that these conclusions would also apply with college freshmen, when conditions equally favorable to effective prediction are present.

Received January 12, 1950.

References

- Altus, W. D., and Mahler, C. A. The significance of verbal aptitudes in the type of occupation pursued by illiterates. *J. appl. Psychol.*, 1946, 30, 155-160.
- Anderson, E. E., and others (Wilson College Studies in Psychology). A comparison of the Wechsler-Bellevue, Revised Stanford-Binet, and Am. Council on Educ. Tests at the college level. *J. Psychol.*, 1942, 14, 317-326.
- Burt, H. E., and Arps, G. F. Correlation of Army Alpha Intelligence Tests with academic grades in high schools and military academies. *J. appl. Psychol.*, 1920, 4, 289-293.
- Cummings, S. B., Jr., MacPhee, H. M., and Wright, H. F. A rapid method of estimating the IQ's of subnormal white adults. *J. Psychol.*, 1946, 21, 81-89.
- Diamond, S. The Wechsler-Bellevue Intelligence Scales and certain vocational aptitude tests. *J. Psychol.*, 1947, 24, 279-282.
- Eurich, A. C., and Cain, L. R. Prognosis. In Monroe, W. S. (Ed.), *Encyclopedia of educational research*. New York: Macmillan, 1940, pp. 838-850.
- Geil, G. A. A clinically useful abbreviated Wechsler-Bellevue Scale. *J. Psychol.*, 1945, 20, 101-108.
- Goldfarb, W. Adolescent performance in the Wechsler-Bellevue Intelligence Scales and the Revised Stanford-Binet Examination, Form L. *J. educ. Psychol.*, 1944, 35, 503-507.
- Gurvitz, M. S. An alternate short form of the Wechsler-Bellevue Test. *Amer. J. Orthopsychiat.*, 1945, 15, 727-733.
- Hunt, W. A., Klebanoff, S. G., Mensh, I. N., and Williams, M. The validity of some abbreviated individual intelligence scales. *J. consult. Psychol.*, 1948, 12, 48-52.
- Hunt, W. A., French, E. G., Klebanoff, S. G., Mensh, I. N., and Williams, M. The clinical possibilities of an abbreviated individual intelligence test. *J. consult. Psychol.*, 1948, 12, 171-173.
- Hunt, W. A., French, E. G., Klebanoff, S. G., Mensh, I. N., and Williams, M. Further standardization of the CVS Individual Intelligence Scale. *J. consult. Psychol.*, 1948, 12, 355-359.
- Hunt, W. A., and French, E. G. Some abbreviated individual intelligence scales containing non-verbal items. *J. consult. Psychol.*, 1949, 13, 119-123.
- Hunt, W. A., and French, E. G. A second fifteen-word vocabulary test for use with abbreviated intelligence scales. *J. consult. Psychol.*, 1949, 13, 124-126.
- Kriegman, G., and Hansen, F. W. VIBS: A short form of the Wechsler-Bellevue Intelligence Scale. *J. clin. Psychol.*, 1947, 3, 209-216.
- Kutash, S. B. A comparison of the Wechsler-Bellevue and the Revised Stanford-Binet Scales for adult defective delinquents. *Psychiat. Quart.*, 1945, 19, 677-685.
- Lewinski, R. J. The psychometric pattern: I. Anxiety neurosis. *J. clin. Psychol.*, 1945, 1, 214-221.
- Patterson, C. H. A comparison of various 'short forms' of the Wechsler-Bellevue Scale. *J. consult. Psychol.*, 1946, 10, 260-267.
- Patterson, C. H. A further study of two short forms of the Wechsler-Bellevue Scale. *J. consult. Psychol.*, 1948, 12, 147-152.
- Patterson, C. H. *Using the Wechsler-Bellevue Intelligence Scales*. In-Service Training Section, Personnel Div., VA Regional Office, Minneapolis, Minn., 1948.
- Rabin, A. I. A short form of the Wechsler-Bellevue Test. *J. appl. Psychol.*, 1943, 27, 320-324.
- Sartain, A. O. A comparison of the New Revised Stanford-Binet, the Bellevue Scale, and certain group tests of intelligence. *J. soc. Psychol.*, 1946, 23, 237-239.
- Springer, N. N. A short form of the Wechsler-Bellevue Intelligence Test as applied to the naval personnel. *Am. J. Orthopsychiat.*, 1946, 16, 341-344.
- Thorndike, R. L. Two screening tests of verbal intelligence. *J. appl. Psychol.*, 1942, 26, 128-135.
- Wechsler, D. *The measurement of adult intelligence*. Baltimore, Williams and Wilkins Co., 1944.

The Minnesota Clerical Test: Sex Differences and Norms for College Groups

Olga E. de Cillis Engelhardt

Department of Psychology, The University of Connecticut

The Minnesota Clerical Test is extensively used in business and industry (5), in vocational advisement (3), as a laboratory exercise in the introductory course in psychology (6), and to predict achievement in selected college courses (4). The manual for this test is replete with 54 different sets of norms, none of which, however, are for college students.

In connection with a larger research project being conducted by the School of Business Administration at The University of Connecticut, the author had occasion¹ to admin-

Paterson (1) and Schneider and Paterson (7) for adults in the general population.

Subjects and Procedure

The subjects of this investigation were 512 junior and senior undergraduate students at The University of Connecticut. Of these 141 male students and 101 female students were enrolled in the College of Arts and Sciences in advanced courses in English, government, psy-

Table 1

Norms for Eastern Arts and Sciences College Students

Centile	Women		Men	
	Test 1 Numbers	Test 2 Names	Test 1 Numbers	Test 2 Names
99	187	189	166	172
90	164	177	147	151
80	148	164	137	144
70	139	155	131	136
60	133	144	124	128
50	128	140	116	120
40	121	131	109	115
30	115	123	105	104
20	107	108	98	97
10	100	93	88	84
1	80	77	75	67
N	101	101	141	141

Table 2
Norms for Eastern Business Administration College Students *

Centile	Men	
	Test 1 Numbers	Test 2 Names
99	185	192
90	153	160
80	136	146
70	130	138
60	124	130
50	118	123
40	115	117
30	110	110
20	105	103
10	95	95
1	80	77
N	270	270

* There is not a sufficient number of women enrolled in the School of Business Administration to warrant the publication of norms.

ister the Minnesota Clerical Test to large groups of college students. The purpose of this brief paper is two-fold: (1) to present a needed set of norms for college students, and (2) to confirm for college students the finding of a sex difference in performance previously reported by Loevinger (2), and Thatcher (2) for grade school children, by Schneider (2) for high school children, and by Andrew and

¹ Graciously provided by Dean Lawrence J. Ackerman.

chology, and zoology. The Arts and Science men ranged in age from 20 to 28; the median age was 21. The women ranged in age from 19 to 26; the median age was 20. The School of Business Administration men consisted of 270 enrolled in a course in industrial psychology. They were primarily industry and marketing majors. These students ranged in age from 20 to 32; the median age of this group was 24. The Minnesota Clerical Test was administered

Table 3

A Comparison of the Medians for the College Groups and Those Reported in the Minnesota Clerical Test Manual *

Groups	Women		Men	
	Test 1 Numbers	Test 2 Names	Test 1 Numbers	Test 2 Names
Arts and Science Students	128	140	116	120
Business Administration Students	—	—	118	123
Eastern Clerical Applicants	128	132	114	121
Southern Clerical Applicants	103	102	—	—
Adults Gainfully Occupied	109	111	83	78
Employed Clerical Workers	144	152	135	126

* (2, p. 8.)

Table 4

Means and Standard Deviations of the Minnesota Clerical Test Scores

Test	N	Mean Raw Scores	Standard Deviations
Numbers			
Arts and Sciences Women	101	128.63	25.02
Arts and Sciences Men	141	117.32	21.95
Business Administration Men	270	121.52	21.74
Names			
Arts and Sciences Women	101	136.70	29.35
Arts and Sciences Men	141	119.66	25.45
Business Administration Men	270	125.44	25.17

Table 5

A Comparison of the Means and Standard Deviations of the College Groups with the General Population and the General Clerical Population *

Population and the General

Test 1, Numbers							
	Gen. Population		Gen. Clerical		A.S. College		B.A.
	Men	Women	Men	Women	Men	Women	Men
M					117.32	128.63	121.52
σ	83.1	113.1	136.1	138.6	21.95	25.02	21.74
N	29.2	30.8	27.9	26.3	141	101	270
	491	229	41	59			
Test 2, Names							
	Gen. Population		Gen. Clerical		A.S. College		B.A.
	Men	Women	Men	Women	Men	Women	Men
M					119.66	136.70	125.44
σ	77.6	111.3	130.0	144.4	25.45	29.35	25.17
N	34.3	36.5	29.4	28.9	141	101	270
	491	229	41	59			

* (2, p. 304.)

during regular class periods during the academic year 1949-1950.

Results

Norms. The normative data for the college population are presented in Tables 1 and 2. A comparison of the medians obtained in the present study with those reported in the manual (2, p. 8) is made in Table 3. The college population medians for both tests are above those reported for adults gainfully employed and southern clerical applicants, but are below the medians for employed clerical workers. The medians of this study are most comparable to those reported for eastern clerical applicants.

Mean Raw Scores and Standard Deviations. The means and standard deviations for both tests for the college population are presented in Table 4. A comparison of the general population and general clerical population with the sample of this study is made in Table 5. College men and women obtain a higher average score than adults in the general population but a lower score than the general clerical population. With the exception of the standard deviation obtained in the present study, for women on test 2, all other measures of dispersion are smaller for the college group

than for the comparison groups (Table 5). This is to be expected since the college population is more homogeneous.

Differences between Men and Women. The differences and their significances for both tests are shown in Table 6. All but one *t* value is significant at beyond the .1% level. The difference between Business Administration men and Arts and Sciences women is significant between the 1% and 2% level of confidence. The sex differences obtained in this study, although highly significant, are not as large as those reported in the literature (1, 7). The significant differences between college men and women found in this study, however, substantiate the earlier findings.

Summary

One hundred and forty-one male and 101 female undergraduate students in the college of Arts and Sciences and 270 males in the School of Business Administration at The University of Connecticut took the Minnesota Clerical Test. Sets of norms for these groups are presented. The sex difference as indicated by other studies on grade and high school children and adults is extended to include college students.

Received August 14, 1950.
Early publication.

Table 6

Differences between College Men and Women on the Minnesota Clerical Test

Test	Means	Mean Diff.	Diff.	<i>t</i>
Numbers				
A.S. Men	117.32	11.31	3.10	3.65**
A.S. Women	128.63			
B.A. Men	121.52	7.11	2.82	2.52*
A.S. Women	128.63			
Names				
A.S. Men	119.66	17.04	3.62	4.71**
A.S. Women	136.70			
B.A. Men	125.44	11.26	3.30	3.41**
A.S. Women	136.70			

* Significant between the one and two per cent levels of confidence.

** Significant beyond the one per cent level of confidence.

References

1. Andrew, D. M., and Paterson, D. G. Measured characteristics of clerical workers. *Bull. Empl. Stab. Res. Inst.*, Univ. of Minnesota, 1934, 3, No. 1.
2. Andrew, D. M., and Paterson, D. G. *Minnesota Clerical Test: 1946 revised manual for the Minnesota Vocational Test for Clerical Workers*. New York: The Psychological Corp., 1946.
3. Baker, G., and Peatman, J. G. Tests used in Veterans Administration Advise ment Units. *Amer. Psychologist*, 1947, 2, 99-102.
4. Barrett, D. M. Prediction of achievement in type-writing and stenography in a liberal arts college. *J. appl. Psychol.*, 1946, 30, 624-630.
5. Bennett, G. K., and Wesman, A. G. Industrial test norms for a southern plant population. *J. appl. Psychol.*, 1947, 31, 241-247.
6. Orbison, W. D., Bousfield, W. A., and de Cillis, O. E. *Group demonstrations and exercises in psychology*. Edwards Brothers, Inc., Michigan, 1950, 196.
7. Schneider, G. G., and Paterson, D. G. Sex differences in clerical aptitude. *J. educ. Psychol.*, 1942, 33, 303-309.

An IBM Card Profile for the Strong Vocational Interest Blank

Wilbur L. Layton

Student Counseling Bureau, University of Minnesota

In the winter of 1949 the Strong Vocational Interest Blank was administered to approximately 7,500 high school seniors in the state of Minnesota. This testing was a part of the state-wide high school testing program sponsored by the Association of Minnesota Colleges. Such large scale use of the Strong Blank was made possible by the cooperation of Mr. Elmer Hanks of Engineers Northwest.¹

cupational scales printed on the back. In addition, the standard scores were coded so that frequency distributions and hence, statistics, could be obtained for each occupational scale.

The code system developed for each occupational scale on the Hanks report forms for men and women is shown in Table 1.

It will be seen that the code numbers 3 and 4, 5 and 6 can be combined to form class intervals of ten standard score units each. This gives seven class intervals of ten standard

HANKES REPORT FORM FOR—
STRONG VOCATIONAL INTEREST TEST - MEN

OCCUPATION	C	C +	B -	B	B +	A
I ARTIST						
PSYCHOLOGIST						
ARCHITECT						
PHYSICIAN						
OSTEOPATH						
DENTIST						
II MATHEMATICIAN						
ENGINEER						
CHEMIST						
III PERSONNEL MGR.						
IV FARMER						
AVIATOR						
CARPENTER						
PRINTER						
TYPEWRITER						
SALESMAN						
POLICEMAN						
FOREST SERVICEMAN						
V TRCA PNT. DIRECTOR						
PERSONNEL DIRECTOR						
FIN. ADMINISTRATOR						
V.M.C.A. SEC.						
SEC. SCH. N.E. TEACHER						
CITY SCHOOL SGT.						
MUNICIPAL						
VI MUSICIAN						
VII C.P.A.						
VIII ACCOUNTANT						
OFFICE MAN						
PURCHASING AGENT						
BANKER						
MODIST						
IX SALES MANAGER						
REAL EST. SALESMAN						
TYPE. MGR. SALESMAN						
X ADVERTISING MAN						
LAWYER						
EDITOR-JOURNALIST						
XI PRES. MFG. COORDIN.						
INTEREST MATURITY						
OCCUPAT. LEVEL						
MASCULINITY TEND.						
STANDARD SCORE						

FIG. 1. Sample card for reporting Strong Vocational Interest results for men.

In order to eliminate the expense of manual duplication and filing of the profiles resulting from this program, the author designed two cards, one for men, one for women.

The cards were designed in a way such that holes punched in the front of the cards would indicate the letter grades on the various oc-

¹ Strong, E. K., Jr., and Hanks, E. J. A note on the Hanks Test Scoring Machine. *J. appl. Psychol.*, 1947, 31, 212-214.

Table 1

Transmutation Table for Occupational Scales

IBM Code Number	Standard Score	Strong Letter Rating
9	65-74	A
8	55-64	A
7	45-54	A
6	40-44	B+
5	35-39	B
4	30-34	B-
3	25-29	C+
2	15-24	C
1	5-14	C
0	-10-+4	C

Table 2

Transmutation Table for O. L., I. M. and M. F. Scales on Men's Form

IBM Code Number	Standard Score
9	70-79
8	65-69
7	60-65
6	55-59
5	50-54
4	45-49
3	40-44
2	30-39
1	20-29
0	10-19

Table 3
Transmutation Table for M. F. Scale on
Women's Form

IBM Code Number	Standard Score
9	20-29
8	30-34
7	35-39
6	40-44
5	45-49
4	50-54
3	55-59
2	60-64
1	65-69
0	70-79

score units each and one class interval of 15 standard score units. Since standard scores below a minus five occur very rarely only a little error is introduced by assuming that one has eight class intervals of ten units each so that means, standard deviations and other

statistics can be computed from the distribution obtained from the IBM cards.

The code for Occupational Level, Interest Maturity, and Masculinity-Femininity scales on the form for men is shown in Table 2.

Code numbers 3 and 4, 5 and 6, and 7 and 8 can be combined to give a total of seven class intervals of ten standard scores each.

The code used for the Masculinity-Femininity scale on the form for women is shown in Table 3.

Code numbers 1 and 2, 3 and 4, 5 and 6, and 7 and 8 can be combined to give a total of six class intervals of ten standard scores each.

This compromise between precise statistical coding and that needed for counseling seems adequate for most purposes. In the near future a revision of these forms will include several new scales which Hankes is planning to add to his present services. Shaded or "chance" areas on the various scales will also be incorporated in the revision.

Received February 16, 1950.

The Super-Roper Technique as a Measure of Interest in Nursing

Leslie Navran

Stanford University

In 1939, a systematic program of research was begun at Clark University under the direction of Donald Super (1, 4, 5, 7, 8) which was designed to explore the possibilities of an objective technique of interest measurement which involved a motion picture presentation of interests. Super's rationale was that a real interest in a vocational field would be manifested via attention to the film, and the closeness of attention and hence the amount of interest could be measured by a post-film test for recognition of what had been seen.

This paper is concerned primarily with the initial study, reported by Super and Roper (8). They showed a strip film (6) depicting the activities involved in nursing, as well as giving information about opportunities in the field, to 35 nurses and 111 high school senior girls, 33 of whom expressed a desire to go into nursing. The subjects were not told in advance of the post-film test. The Strong Vocational Interest Blank and the Otis S-A Test of Mental Ability (Higher Form, A) were also administered. For both nurses and students, the correlations between the Super Test of Interest and each of the other tests were zero. The interpretation suggested was that the Super Test was measuring something other than intelligence or subjective interest patterns. The authors argued that it was logical, therefore, to assume that the Super was a measure of interest in the occupation itself. The odd-even reliability of the test, corrected, was .83. Further analysis revealed that Super's test differentiated the nurses from the students, and within the latter group, those who wanted to be nurses from those who did not.

Older (4) has summarized the studies which were subsequently carried out. It need only be noted here that Bernstein (1) reported a correlation in the low .30's between the Super-type test for the metal trades and the Otis, when subjects expected a post-film test, and that Super and Haddad (7) found that Super-type test results were not significantly correlat-

ed with familiarity with or information about the occupations covered by the film. Older's (4) study has been the last reported to date. His results, based on a film which covered not one, but a range of occupations, were similar, in many respects, to those reported by Super and Roper.

The Problem

While Super's goal was to achieve a test which would be occupationally comprehensive, an equally valuable goal would be the further development of such a technique for a given specific field. Roper's (5) results raised the questions of: (1) the usefulness of such a technique in the selection of students for nursing; and (2) the relation between a high interest "in the work itself" and achievement. She made an attempt to investigate these questions, but the negative results were inconclusive, since the criteria were considered defective, and the sample included only *successful* trainees. This study was an attempt to make a more thorough investigation of the above questions. The focus of interest was not on the Super Test of Interest in Nursing per se, but in the technique and its potentialities.

Procedure

The subjects were 70 girls entering nurses' training in September 1947. Forty-four were divided among three hospitals affiliated with San Jose State College, and the remaining 26 were enrolled at a hospital affiliated with the Stanford University School of Nursing.

A Personal Data Sheet and an Information Blank, both of which were devised by the writer for this study, were added to Roper's (5) test battery. The Information Blank consisted of 15 items which formed a continuum of nursing knowledge ranging from that which is commonly known to that which is more likely to be known only by nurses. The test was pre-tested on a nursing sample and on a sample of senior high school girls not interested in nursing as a career.

In order to motivate the subjects, the battery was administered as a part of their orientation program. The order of the tests was as follows: Information Blank, film and Super Test, Personal Data Sheet, Otis, and Strong. Except for the new procedures, this was the order used by Roper. All the girls were free from other commitments for at least one hour before testing. Ventilation, visibility of the screen, and other conditions were as optimal as possible.

Two changes were made with respect to the film: (1) one of the film's 40 frames was exposed for 17 seconds instead of the usual 15, to insure its being completely read; and (2) the subjects were instructed to expect a post-film test which would be based on the film. This has been acknowledged by Super as a practical necessity.¹ With these exceptions, the comparison made later with Roper's data is between data obtained under essentially the same conditions.

The scoring of the Otis and Strong was in accordance with their respective manuals. The scoring key for the Super Test was supplied in Roper's (5) thesis, and verified by both Roper and Super.

The following two criteria of achievement were selected: (1) grades in courses taken during the first year of training, and (2) ratings of performance of duties while on the ward. Test results were withheld from the hospital and school administrators until the data relevant to the criteria were obtained, in order to avoid contamination of their judgements.

Results

In the first two years of training, ten girls left the San Jose group of trainees, four to be married and six because of scholastic difficulties. One girl left the Stanford class to be married.

Comparison of the two student groups' Personal Data Sheets yielded these major differences: (1) The Stanford girls were a more homogeneous age group and were two years older in median age. (2) Eighty-seven per cent of the San Jose girls had one year or less of college education, in contrast to the one Stan-

ford girl (four per cent) who had less than two college years. This was a significant difference ($C.R. = 7.0$ p ; $< .001$). (3) Only 52 per cent of the 19 San Jose girls who had some college education had major subjects related to nursing. In contrast, 88 per cent of the Stanford girls were nursing majors ($C.R. = 2.6$; $p < .01$).

The percentage of the girls in each group having work experience and previous social contact with medical persons was approximately the same. For both groups, the decision to enter nursing was based on a matrix of main reasons: a long-term ambition to be a nurse, a desire to help and work with people, an assortment of practical reasons, and a liking for the work. The most frequently given reason was: discussion with, and urging by, friends, family and counselors. However, only 12 of the 70 girls reported having had psychometric guidance.

The Information Blank significantly differentiated all the groups to which it was administered (see Table 1).

Going from the non-nurses to the trainees and nurses, there is a significant increase in the mean scores, a consistent drop in the variability, and a narrowing of the range. These results are what one would expect if the test were valid. The test also significantly discriminated the more highly educated and presumably better informed about nursing Stanford group from the San Jose group ($C.R. = 3.75$; $p < .001$). Although it may be argued that the small number of items did not

Table 1

Summary of Information Blank Scores

	H. S. Seniors (N = 82)	Nursing Trainees (N = 70)	Nurses (N = 51)
Mean	5.7	10.2	12.7
S.D.	2.45	2.2	1.4
Range	1-12	5-15	8-15

Critical Ratios

Nurses $\times$ Trainees	7.8 ($p < .001$)
Trainees $\times$ H. S. Girls	10.7 ($p < .001$)
Nurses $\times$ H. S. Girls	20.8 ($p < .001$)

¹ Personal communication.

Table 2

A Comparison of the Most Similar Components of Roper's and Navran's Sample with Respect to the Otis, Strong, and Super Tests

Test	Roper H. S. Girls Indicating Choice of Nursing (N = 33)	Navran H. S. Graduates Entering Nurses' Training (N = 44)	Critical Ratio
Otis {Mean S.D.	108.9 9.9	110.8 7.95	.90
Strong {Mean S.D.	57.1 6.5	40.1 18.7	5.57 ($p < .001$)
Super {Mean S.D.	51.1 12.0	87.5 13.2	12.59 ($p < .001$)

tap all possible areas of nursing information, the obtained results indicate that these 15 items are a valid sampling of nursing information, which successfully differentiates better informed from less informed groups.

The mean Otis IQ of the Stanford group was 114.3, while that for the San Jose group was 110.8 (C.R. = 2.1; $p < .04$). Only the Strong failed to discriminate the two groups. The Stanford mean was 43.8 and the San Jose mean 40.1.

Super Test. For the purpose of investigating the test's reliability, all 70 subjects were grouped together. The correlation of the odd versus even items for the whole test was .54, raised to .70 by the Spearman-Brown formula. The first four parts of the test correlated .97 with the entire test. This supports Roper's (5) view that Part V is virtually useless. The odd-even reliability coefficient of these four parts was .48, raised to .65. The first three parts, reported by Super and Roper (8) as the most valuable for discriminating nurses from non-nurses, had a reliability of .52, raised to .69.

Comparison of Roper's Results with Those of the Present Study. As the San Jose group was closest to the 33 high school seniors in Roper's sample with respect to educational level, only the scores of these two groups were compared. This comparison is of interest, despite the San

Jose group's being more highly selected (see Table 2).

The two groups were not significantly different in intelligence, as measured by the Otis. However, the two groups differed significantly on both the Strong and Super tests, Roper's sample having a marked edge on the Strong, while the subjects used in this study were superior on the Super.

The Strong Test comparison is clouded by the fact that a new revision of the test was used in this study. Ten items of the older form were cut out in this revision. Accordingly, some of the difference may be an artifact attributable to the Roper group's having a greater number of scoring opportunities. Probably of greater importance is the likelihood that the two groups, being seven years apart, actually differ with respect to the nursing pattern of likes and dislikes.

The Super Test difference may be accounted for by three factors: the extremely low reliability of the test, the fact that the subjects in this study expected the test, and the difference in selection.

Test Intercorrelations. A comparison of the interrelationships among the tests with the intercorrelations obtained by Roper (5) is presented in Table 3.

The correlation between the Super and Otis tests is the only one noticeably changed; the r of .345 is significant. This increase is in accord with the results reported by Bernstein (1) and Older (4).

Table 3

A Comparison of the Obtained Test Intercorrelations with Those Reported by Roper

	Navran (N = 70)	Roper Nurses (N = 35)	Roper H. S. Girls (N = 111)
Super $\times$ Otis	.345 ($p < .004$)	.006	.14
Super $\times$ Strong	.055	-.008	.001
Strong $\times$ Otis	.016	-.04	.00
Information $\times$ Otis	.395 ($p < .004$)	—	—
Information $\times$ Super	.198	—	—
Information $\times$ Strong	.001	—	—

The relation between information and intelligence, as measured by these tests, was positive and significant. The Super and Information Blank showed a positive but insignificant relationship. Evidently, the two tests are measuring essentially different functions. These results agree with those of Haddad and Super (7).

Comparison of Successful and Unsuccessful San Jose Trainees. An analysis of the test performance of the ten girls who left training, as compared to the 34 who remained, revealed no significant differences. However, the evidence indicated that those who left to get married were students as promising as those still in training. The scholastic failures scored lower than the marrying girls on each of the four tests, but only the difference in intelli-

Table 4

The Correlation of Tests in the Battery with the Criterion of Grades

Test	San Jose (N = 34)	Stanford- Lane (N = 25)	Critical Ratio
Otis	.33 ($p < .06$)	.65 ($p < .001$)	1.55
Information	.45 ($p < .01$)	.28	
Super	.20	.12	
Strong	.03	-.04	

gence was significant ($t = 2.32$, 8 d.f.; $p < .05$).

The Relation Between Test Scores and the Criterion of Grades. For each student group, test scores were correlated with the average number of grade points for courses during the first school year (see Table 4). The 12 grades from A to F were equated with numerical values from zero to eleven, and the usual formula for computing grade (honor) points was used. The grade point average was obtained by dividing the total number of grade points by the total number of course units.

Neither interest measure correlated significantly with the criterion. The Otis correlations were low to moderately positive, and it would be the most useful test in the battery for prediction. Only in the San Jose group was the correlation for the Information Blank

Table 5

The Correlation of the Parts of the Super Test with the Criterion of Grades

	San Jose (N = 34)	Stanford-Lane (N = 25)
Part I	.41 ($p < .02$)	.61 ($p < .01$)
Part II	-.20	.10
Part III	-.004	.12
Part IV	.35	.02
Parts I, II, III	-.002	.28
Total Test	.20	.12

significantly positive. This suggests the possibility that with increased intelligence the function of previous information is minimized in mastering course work.

The relationship of the various parts of the Super to the criterion is shown in Table 5.

Only Part I, having 21 multiple-choice items, was significantly related to the criterion in each group. The variability of the correlations of the other parts, both in amount and direction, was evidence of both their unreliability and their lack of relationship to the criterion.

The Relation Between Test Scores and the Criterion of Ratings of Performance of Duty in the Wards. Only the ratings for the Santa Clara County Hospital and Stanford Hospital trainees were suitable for analysis (see Table 6).

The correlations for the Santa Clara group are all lower than the Stanford correlations, but the rank order is the same in both groups. The Super Test correlations are not only low,

Table 6

The Correlation of Tests in the Battery with the Criterion of Ratings of Performance of Ward Duties

Test	Santa Clara County Hospital (N = 11)	Stanford-Lane Hospital (N = 24)
Otis		.37
Strong	.10	-.17
Super	-.32	-.25
Information	-.32	.04
	-.22	

Table 7

The Correlation of the Parts of the Super Test with the Criterion of Ratings of Performance of Ward Duties

	Santa Clara County Hospital (N = 11)	Stanford-Lane Hospital (N = 24)
Part I	-.47	.24
Part II	.07	-.06
Part III	-.76 ($p < .01$)	.06
Part IV	-.25	-.26
Parts I, II, III	-.38	.08
Total Test	-.322	-.25

but negative. Only the Otis is correlated positively with the criterion in both groups. However, none of the correlations are significant.

Table 7 summarizes the correlations between the parts of the Super and the criterion.

The correlations for the Santa Clara group, based on only 11 cases, are not considered as reliable as the Stanford correlations; but neither set of coefficients indicates that the Super is related positively to the criterion.

Discussion

Reliability and Validity. It has been repeatedly demonstrated that the Super technique discriminates groups; but in view of the present results, it is pertinent to question what function is being discriminated.

There are two objections to the hypotheses advanced by Super and Roper (8), and also by Older (4), that it is logical to assume the Super Test is measuring interest in the occupation itself because of its low correlations with the Strong and the Otis. First, the obtained reliability coefficient, corrected, of .70 for the total Super Test, is noticeably lower than the coefficient of .83 previously reported (8). Taken together with the fact that any odd-even reliability coefficient will be too high because the assumption of uncorrelated errors is not met (2), the test's reliability is of such a low order as to cast serious doubt on its validity. Moreover, such low reliability would act to decrease all correlations of the Super

with other tests, making the *true* relationships among them ambiguous. Since its relationship to intelligence is already significant, and that with information of near-significance, the possibility must be entertained that increased reliability for the Super-type test might show these factors to play a more significant role than has been thought.

Secondly, there is a serious danger in the reasoning of the authors cited above: if one is to ascertain what a test measures by eliminating what it does *not* measure, then all the variables which might reasonably be important should be considered. Unless this is done, any conclusion must be somewhat precarious.

Interpretation of Results. The lack of relationship of Super Test scores to each of the criteria cannot be attributed to mechanical imperfections in the assigning of grades and ratings. Space does not permit the extended treatment given this issue elsewhere (3); suffice it to say that inspection of the data revealed the spread of grades in the courses covered to be quite representative. Discrimination was also enhanced by the treatment of the data. With respect to the ratings, the present study is admittedly weak because of the small number of cases. The lack of a uniform rating system for all the San Jose hospitals, as well as an opportunity to instruct the raters, made the loss of so many San Jose subjects unavoidable. Although the small size and relative homogeneity of the Stanford group was regrettable, it is doubtful if many nursing schools have given as much thought and consideration to the adequacy of their rating system. Accordingly, there seems to be no reason to question the *quality* of the ratings for this group.

The Super's low reliability, with its depressant effect on the intercorrelations with the criteria, offers the most parsimonious interpretation. The amount of error in the Super is so great as to make it uncertain just what the test is measuring. Consequently, the obtained results cannot be viewed as reflecting the relationship between objective interest and achievement. The true relation will be open to investigation when two conditions are satisfied: (1) a reliable measuring instrument is made available; and (2) it is *positively* estab-

lished that the technique is measuring objective interest.

In connection with the latter point, attention should be given to the possibility that something else other than interest is being measured. As presently constituted, two-thirds of the items in the Super Test of Interest in Nursing are descriptions of pictures which may or may not have been seen in the film. All the items in Older's (4) film were of this type. Thus, the tests favor those subjects who best recognize the word descriptions corresponding to the images they have of the pictures they have seen. Super et al. make the assumption that this ability stems primarily from interest. It might prove enlightening to test the relation between a laboratory test of visual imagery and performance on the Super-type test. The relation of such performance to intelligence might be more clearly and adequately defined if the Otis were replaced by a test which measures specific aptitudes or abilities. It might also be that, in this study, the Super technique is not discriminating between more and less interested people, but, instead, between groups whose differential amounts of exposure to medical stimuli make their perception of nursing scenes proportionately more meaningful. This differential might be because of interest, but it might also be a matter of circumstance.

Implications for Future Research with the Super Technique. Future research should be concerned first with positive efforts to establish what the technique measures. The factors cited above, as well as other possibilities, should be investigated. Coordinated with this should be attempts to revise the existent Super-type tests so as to increase their reliability. In making these revisions, the results of this study indicate that multiple-choice items, used in Part I of the Nursing Test, should be given greater weight; i.e., increased in number. Not only did Part I correlate significantly with grades, but this is the most reliable type of item employed by Super. The establishment of interest as the variable being

measured and the construction of a reliable instrument would then permit the investigation of the relation of objective interest to achievement.

Summary

1. The relation of scores on the Super Test of Interest in Nursing to the criteria of grades and ratings of performance of duties on the ward was not significantly above a chance level.

2. In view of the low reliability of the Super Test, the above results are ambiguous. The variable being measured may or may not be more related to the criteria. Such a determination must await a more reliable instrument.

3. It will be worthwhile to establish more positively that interest is the variable being measured. The uncertainties resulting from the technical defects of this and previous studies make desirable the further investigation of such other factors as experience and special aptitudes.

Received January 6, 1950.

References

1. Bernstein, Benjamin. *A test of interest in the metal trades*. Unpublished Master's thesis, Clark University, 1940.
2. McNemar, Q. *Psychological statistics*. New York: John Wiley and Sons, Inc., 1949.
3. Navran, L. *The relation of interest to achievement in nursing*. Unpublished Master's thesis, Stanford University, 1949.
4. Older, H. J. An objective test of vocational interest. *J. appl. Psychol.*, 1944, 28, 99-108.
5. Roper, Sylvia A. *A test of interest in nursing*. Unpublished Master's thesis, Clark University, 1940.
6. Strip film: *Nursing as a career*. Chicago: Society for Visual Education, Inc.
7. Super, D. E., and Haddad, W. E. The effect of familiarity with an occupational field on a recognition test of vocational interest. *J. educ. Psychol.*, 1943, 34, 103-109.
8. Super, D. E., and Roper, Sylvia A. An objective technique for testing vocational interests. *J. appl. Psychol.*, 1941, 25, 487-498.

Acuity Differences between the Two Eyes and Job Performance

Newell C. Kephart and Joseph M. Mason

Occupational Research Center, Purdue University

The importance of vision in industry is widely recognized, and the relationships between certain visual skills and numerous aspects of job performance have been emphasized. Investigations by Kephart (3, 4), Stump (6), and Tiffin (7, 8, 9, 10) have convincingly demonstrated that visual skills are related to job performance, for they have found very marked relationships between these skills and hourly production, rate of labor turnover, accident-proneness, supervisors' ratings, and other acceptable industrial criteria. In general, these and numerous other studies had been concerned with individual jobs, and not with a variety of industrial jobs to determine existing basic relationships of general applicability.

To investigate these general relationships, McCormick (5) studied a group of approximately 5,500 employees on 92 jobs in a number of different industrial plants. From his studies he concluded that the visual acuity requirements for acceptable performance on these 92 jobs seemed to be general and relative, and that a moderate minimum of far and near acuity in both eyes was especially pertinent to satisfactory performance. Beyond such a minimum the percentage of high criterion employees increased continuously with increase in acuity level, but at a more gradual rate. The relationship of job success to worse eye acuity was also studied, and he reports that the increase in probability of job success with increases in worse eye acuity is relatively constant for the entire range, both for near and far acuity.

McCormick, as reported, studied only the relationships between job performance and both eye acuity, near and far, and between job performance and worse eye acuity, near and far. Nowhere in the published literature on vision studies, are we able to find an analysis of the visual acuity differences between the two eyes, and the relationship of these differences to job performance. The present investigation is directed toward that objective.

Statement of the Problem

For a number of years, the Occupational Research Center (11) of the Division of Education and Applied Psychology, Purdue University, has been actively engaged in the development of vision-test profiles. In order to assign employees to specific jobs in which their chances of success are greatest, it is necessary to evaluate the different visual requirements of each individual job and to measure in some adequate way the visual skills of the applicant. Industrial and business organizations, subscribing to the vision service,¹ use the profiles developed in the Occupational Research Center to identify individuals whose vision does not meet the general visual demands of their jobs and who could possibly benefit from professional eye care. The profiles are also used as an aid in employee selection.

One question, unanswered by any of the previous vision studies, has to do with the relationship, if any, between job performance and acuity differences between the two eyes. In practically all vision tests, both eye acuity is measured. Then with first one eye occluded and then the other, measures of left and right eye acuities are determined. It has been observed that in a large percentage of individuals, one eye will be considerably better than the other. Is the "one-eyed" individual handicapped in an industrial job? Or does the worker with one very good eye and one very weak eye perform on a comparable level with the worker who has two very good eyes?

Thus, the present study is an extension of the research of McCormick (5), on the relationship between visual skills and job performance. The specific objectives of the investigation are:

1. An analysis of the over-all relationship between better eye acuity, near and far, and job performance of employees on a number of different industrial jobs.

¹ Industrial Vision Service of the Bausch and Lomb Optical Company, Rochester, New York.

2. An analysis of the over-all relationship between acuity differences in the two eyes and job performance.

3. After sub-dividing the total group into a number of sub-groups, on the basis of better eye acuity, an analysis of the acuity differences between the two eyes, as related to job performance, for each of these better eye acuity levels.

Procedures

Basic Data Used. The basic data used for this investigation were Ortho-Rater² test results, and job performance measures available in punched-card form in the Occupational Research Center. The data covered approximately 5,600 employees on 92 jobs in a number of different industrial establishments. With a few exceptions, they are basically the same data used by McCormick (5) for his investigation.

Vision Tests. Of the twelve vision tests included in the Ortho-Rater battery, only four were given attention in this investigation; these were: Near test (given at the optical equivalent of 13 inches), acuity, right eye, acuity, left eye; and far test (given at the optical equivalent of 26 feet), acuity, right eye, and acuity, left eye.

The better eye scores, whether left or right, for each individual, were used directly. By comparing the near acuity test scores of the left and right eyes for each individual, the one with the higher score was considered to be the better. The far acuity test scores for the left and right eyes of each individual were similarly compared and the better eye determined. Where both left and right eye had the same acuity score, the treatment was the same; for such an individual, the better eye and worse eye scores would be identical.

The difference between the better eye score and the worse eye score, for the near acuity tests, was then determined—a simple process of subtracting the one score from the other. A similar computation was made for all far acuity better-eye and worse-eye scores. The four measures of visual acuity that were actually used in the investigation were then: near acuity, better eye; near acuity, difference between better eye and worse eye; far acuity, better eye; and far acuity, difference between better eye and worse eye.

In the Ortho-Rater acuity tests, the range of possible scores is from 0 to 15. These scores can be readily converted to other measures of visual acuity, such as the Snellen notation or the AMA acuity ratings, as explained in the manual of standard practice for the Ortho-

Rater (12). With the large number of test results available for this study, the distribution of better eye scores covered the entire range, although at both the lower and upper levels, the number of results included were relatively small. The greatest difference between better eye and worse eye acuities for any individual was thirteen score points.

Criteria of Job Performance. From one industrial plant to another, and from one job to another, the criteria of job performance varied considerably. In some establishments, the criteria were based on such factors as hourly production and earnings; in other plants, the measures of job performance were based upon supervisors' ratings. In general, however, and regardless of the basis of the criteria, the measurements were reduced to numerical categories, and these were subdivided into two groups, identified as the "high criterion group" employees and the "low criterion group" employees. Occasionally where there was an odd number of numerical categories, it was obviously impossible to subdivide the middle category into the "high and low criterion" groups, and it became necessary to omit these employees from consideration in the analysis. These omissions account for the apparent discrepancy in numbers of test results analyzed; from the initial group of approximately 5,600 employees, about 5,100 could be definitely assigned to one of the two criterion groups.

Data on the Relationship between Better Eye Acuity and Job Performance. In analyzing the over-all relationship between better eye acuity and job performance, summaries of the test scores and the measures of job performance were derived from the basic data. For near acuity, better eye, the per cent of all employees at a given score level who were high criterion group employees was determined. The percentages were computed for each score level for the better eye, from 0 to 15. Similar per cents were determined for each score level for far acuity, better eye. Regarding this data as only one sample out of a larger population having practically the same properties as the ones obtained, it seemed desirable to obtain "smoothed" percentages, thus minimizing irregularities in the data. This was done by a system of running averages. Guilford (2) illustrates this running-average method: for a given score, the two immediately neighboring scores are averaged with the given score, which is allowed to carry double-weight. McCormick (5), in smoothing his data, assigned equal weight to all three scores, and in this way determined the average. Both of these methods of "smoothing" were applied to a part of the data, and gave essentially the same results, so it was decided to use the latter method. The "smoothed" per cent determined for a

² The Ortho-Rater is a stereoscopic-type visual testing instrument manufactured by the Bausch and Lomb Optical Company, Rochester, N. Y.

given score was the quotient of the sum of the frequencies of all (both high and low criterion groups) employees for the corresponding three score levels. "Smoothed" percentages were thus determined for better eye score levels, and the resulting per cents subsequently plotted.

Data on the Over-All Relationship between Acuity Differences in the Two Eyes and Job Performance. To analyze the over-all relationship between acuity differences between the two eyes and job performance, it was necessary to derive the score difference for each individual employee from the basic data of acuity scores for the two eyes. The electric punched-card tabulation equipment available in the Occupational Research Center greatly facilitated this phase of the investigation. The approximately 5,600 cards were sorted as to better eye acuity, and then in a second sorting, according to worse eye. Then the difference between the score of the better eye and the score of the worse eye was punched on the card.

To study the over-all relationship of acuity differences between the two eyes, employees were grouped on the basis of this difference. Thus, all those showing a difference of three score points between the two eyes were placed in one group; such a category would then include an individual scoring 3 on the better eye and 0 on the worse eye, as well as an employee scoring 15 on the better eye and 12 on the worse eye, and all others where the score difference between the two eyes was three points. In similar manner, all employee test results were categorized according to the acuity differences between the two eyes. The per cent of all employees at each difference level who were high criterion group employees was determined. These percentages were computed for each acuity difference level both for the far acuity tests and the near acuity tests. To minimize the irregularities in the data due to sampling, "smoothed" percentages were computed for each difference level in the manner described above. Subsequently, these "smoothed" percentages were plotted to furnish a visual picture of the over-all relationship between acuity differences in the two eyes and job performance.

Data on the Relationship between Acuity Differences in the Two Eyes and Job Performance for Each Better Eye Acuity Level. The test scores and measures of job performance derived from the basic data were divided into 15 sub-groups on the basis of better eye acuity score. The employees represented in any given sub-group then had the same acuity score for the better eye, whereas the acuity scores for the worse eye would vary anywhere from the given acuity score of the better eye down to, and including, a score of 0. This finer breakdown into better eye acuity levels was necessary in order to

study the relationship between acuity differences in the two eyes and job performance at each of these levels. For each sub-group, with a constant better eye acuity score, the per cent of the total at each difference level who were high criterion group employees was determined. Differences between the two eyes at each better eye acuity level were treated for the far tests as well as for the near tests, and the percentages of high criterion employees for each group were determined. "Smoothed" percentages were then calculated for all of these sub-groups in the manner described previously. To present a visual picture of the relationship between acuity differences in the two eyes and job performance, these "smoothed" percentages for each better eye acuity level were subsequently plotted.

Statistical Test of Homogeneity of Sample Applied to Each Sub-Group. Fisher (1) demonstrates a procedure for testing the homogeneity of a sample in a $2 \times n'$ classification which is applicable to the data used in this investigation. This test of homogeneity was applied to all sub-groups, both for the near and far tests, where the total N was sufficiently large to make the test applicable. This resulted in an elimination of only those sub-groups where the better eye acuity score was 3 or less. For all other sub-groups where the total for any given difference level was less than 15, the values were combined with the succeeding difference groups to bring the total to the required 15. In determining the proportions, calculations were carried to seven decimal places, as Fisher states (1: 91) that using five decimal places the value of χ^2 is not quite correct in the second decimal, and to avoid doubts as to the precision of calculation, two more places are desirable.

All χ^2 values determined after applying this test of homogeneity were interpreted in terms of P from the values given in Fisher's Tables (1: 104-105), and subsequently each χ^2 value was compared to the .05 probability values as given in the Tables. This comparison was made because Fisher states (1: 82): "We shall not often be astray if we draw a conventional line at .05, and consider that higher values of χ^2 indicate a real discrepancy."

Results

Relationship between Better Eye Acuity and Job Performance. The relationship existing between better eye acuity and job performance is reflected by the percentage of individuals at each score level who were in the high criterion group of employees. The basic relationship represented by these percentages is shown

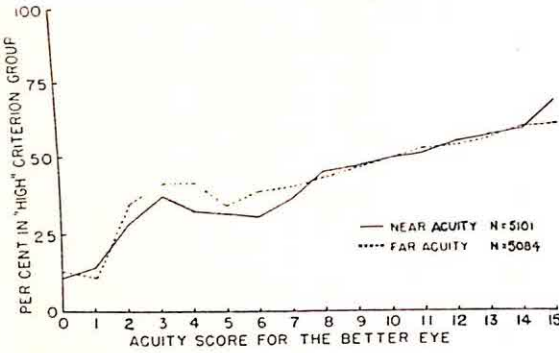


FIG. 1. Relationship between better eye acuity and job performance.

graphically in Figure 1, which presents the relationship for all of the 92 jobs combined, both at the near and far acuity distances.

Both the far and near acuity, better eye, trend lines show a relatively marked increase in the percentage of high criterion group employees at each successive better eye score value. Examination of Figure 1 shows that for all 92 jobs combined there is a general increase in the proportion of individuals who were high criterion group employees at successively higher scores, and that the increase is quite persistent throughout the entire range of scores, though perhaps a little more marked in the lower score range. The few irregularities in the general trend line are probably due to sampling errors, for even though the total N is large, the number of employees scoring in the 0 to 3 and 13 to 15 score levels are relatively small. The general trend presented here graphically would indicate that at least a moderate degree of visual acuity in one of the two eyes, both at near and far distance, is necessary for satisfactory performance on the types of industrial jobs covered by this investigation. Further, it is indicated that the greater the degree of visual acuity in the better eye the greater the probability of success for the employee on these jobs.

The general relationship existing between the better eye acuity and job performance, as indicated in Figure 1, is similar to the relationship determined between both eye acuity and job performance (5). This is true both for the near and far acuity distances.

Over-All Relationship between Acuity Differen-

ces in the Two Eyes and Job Performance. The relationship existing between the visual acuity differences in the two eyes and job performance is reflected by the percentage of employees at each difference level who were in the high criterion group of workers. The basic relationship represented by these percentages is shown graphically in Figure 2. In this figure is presented the relationship for all of the 92 jobs combined, both at the near and far acuity distances.

The acuity differences represented cover the entire range of acuity differences between the better eye and the worse eye. Thus, at any given score difference on the base line, the value plotted directly above it represents the percentage of the total group having that acuity difference who were in the high criterion group of employees. For example, above the point labeled "7" on the base line, are represented all employees showing a difference of 7 score points between the better eye acuity score and the worse eye acuity score; it would include those who scored 7 on the better eye and 0 on the worse eye, as well as those who scored 14 on the better eye and 7 on the worse eye, and all others whose acuity difference between the better and worse eyes was 7 score points.

For acuity differences between the two eyes, there appears to be no relationship between these differences and job performance, indicating that with greater differences between the two eyes (both near and far), the probabilities of satisfactory performance on industrial jobs of the types investigated here are approximate-

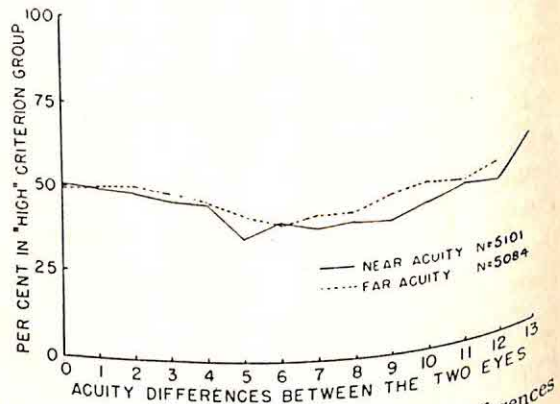


FIG. 2. Over-all relationship between acuity differences in the two eyes and job performance.

ly the same as the probabilities of satisfactory performance where the two eyes are equally good. The general trend presented here graphically indicates the over-all relationship for acuity differences between the two eyes and job performance, and suggests that such acuity differences do not, at any difference level, significantly affect the probabilities of satisfactory performance for the combined group of jobs.

Relationship between Acuity Differences in the Two Eyes and Job Performance for Each Better Eye Acuity Level. For a given better eye acuity, the relationship existing between acuity differences in the two eyes and job performance is reflected by the percentage of employees at each difference level who were in the high criterion employee group. The basic relationships are presented graphically for better eye acuity scores 3 to 15 in Figures 3 to 6, inclusive.³ These figures represent the relationships for all of the 92 jobs combined, both at the near and far acuity distances.

At any given better eye acuity, the apparent lack of relationship from the low acuity differences to the high acuity differences indicates that such differences do not significantly affect the probabilities of satisfactory performance for these jobs. Where the acuity difference between the two eyes is large, the probability of satisfactory performance on these jobs is approximately the same as the probability of satisfactory performance where the difference is small. This general relationship holds for all better eye acuity levels, both near and far.

Results of Statistical Test of Homogeneity of Samples. Fisher's test of homogeneity of a sample in a $2 \times n'$ classification, applied to each sub-group for each better eye acuity score, substantiates the basic relationships indicated in the graphic presentations. None of the proportions of high criterion employees within a given better eye acuity sub-group deviates significantly from the theoretical proportion for the group as a whole. The χ^2

values obtained, the P values interpolated from Fisher's Tables (1:104-105), and the χ^2 values for a P of .05 are presented in Tables 1 and 2,⁴ for far and near acuity distances. Fisher states that if P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested, or that the observed deviations could not have occurred by chance. We have therefore, accepted his suggestion and hold that only where P is smaller than .05, is there sufficient indication of a significant deviation. None of the P values for any better eye acuity level, near or far, is as small as .05; the range for the 24 sub-groups is from .09 to .82, with 23 of the 24 falling within the .1 to .9 limits mentioned above.

This statistical test then confirms the observed relationship between acuity differences of the two eyes and job performance, at constant better eye acuity levels. For the 92 industrial jobs combined, as studied in this investigation, acuity differences are not related to probabilities of satisfactory job performance.

Summary

An investigation was made of the existing relationship between certain visual skills and job performance on a variety of industrial jobs. Ortho-Rater scores of approximately 5,600 employees on 92 different industrial jobs comprised the basic data. Using various measures of job performance as criteria, about 5,100 of these employees were identified as either high criterion group employees or low criterion group employees. Four measures of visual acuity were given particular attention and their relationship to job performance analyzed. These four acuity measures were: better eye acuity, near; better eye acuity, far; acuity differences between the two eyes, near; and acuity differences between the two eyes, far. Acuity differences between the two eyes were analyzed both for an over-all relationship to job performance, and for each better eye acuity level.

The following conclusions are drawn from this investigation:

1. For jobs of the type studied, a moderate minimum of visual acuity in one eye or the

³ To reduce printing costs, Figures 3, 4, 5, and 6, and Tables 1 and 2 have been deposited with the American Documentation Institute and may be ordered as Document 2870 from American Documentation Institute, 1719 N Street, N.W., Washington 6, D. C., remitting \$0.50 for microfilm (images 1 inch high on standard 35 mm. motion picture film) or \$0.60 for photocopies (6 x 8 inches) readable without optical aid.

⁴ See footnote 3.

other is necessary for satisfactory job performance.

2. Beyond this minimum, probability of satisfactory performance increases quite constantly with greater degrees of acuity in the better eye.

3. In the over-all relationship between acuity differences in the two eyes and job performance, such differences are not significantly related to job performance.

4. A horizontal straight-line relationship exists between acuity differences in the two eyes and job performance, so that the probability of satisfactory performance where acuity differences are large is approximately equal to the probability of satisfactory performance where acuity differences are negligible.

5. For any given better eye acuity level, acuity differences in the two eyes do not affect the level of job performance for the types of jobs studied.

This investigation has been primarily concerned with the over-all relationships for the 92 jobs combined. It is possible that on certain industrial jobs that require particular types of visual skills, acuity differences in the two eyes may be related to job performance. An extension of this investigation to employees on particular jobs would be necessary to determine these relationships.

Received January 23, 1950.

References

1. Fisher, R. A. *Statistical methods for research workers*. (4th Ed.) London: Oliver and Boyd, 1932.
2. Guilford, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill Book Company, Inc., 1942.
3. Kephart, N. C. An analysis of professional eye care and industrial efficiency. *Trans. Amer. Acad. Ophthal. Otolaryng.*, 1946, **51**, 166-170.
4. Kephart, N. C. Visual skills and labor turnover. *J. appl. Psychol.*, 1948, **32**, 51-55.
5. McCormick, E. J. *An analysis of visual requirements in industry*. Ph.D. Thesis, Purdue University, 1948.
6. Stump, N. F. Vision tests predict worker capability. *Factory Mgmt. and Maint.*, 1946, **104**, 121-124.
7. Tiffin, J. *Industrial psychology*. (2nd Ed.) New York: Prentice-Hall, Inc., 1947.
8. Tiffin, J. The use of visual data as an aid to increase production and efficiency. *Trans. Amer. Acad. Ophthal. Otolaryng.*, 1944, **49**, 1-4.
9. Tiffin, J., and Greenly, R. J. Employee selection tests for electrical fixture assemblers and radio assemblers. *J. appl. Psychol.*, 1939, **23**, 240-263.
10. Tiffin, J., and Wirt, S. E. The importance of visual skills for adequate job performance. *J. consult. Psychol.*, 1944, **8**, 80-89.
11. Wirt, S. E. Statistical laboratory for vision tests at Purdue University. *J. appl. Psychol.*, 1946, **30**, 354-358.
12. *Standard practice in the administration of the Bausch and Lomb occupational vision tests with the Ortho-Rater*. Bausch and Lomb Optical Company, Rochester, N. Y., 1944.

Point Centering of Signals on an Area *

Adelbert Ford, David Rigler, and Genevieve E. Dugan

Lehigh University

This task relates to the precision with which signals, consisting of white spots or patches, can be located or "tracked" on a dark field, such as that of a radar scope, by means of a device by which an operator can move a point cursor to the center of the signal, using a pantograph lever system giving two degrees of muscular freedom, and experimental amounts of movement reduction.

The problem includes the discovery of the absolute amount of the error, its comparison with the sizes of errors previously reported for scale reading techniques of locating signals, the presentation of any tendencies towards drifts and other variability in efficiency, and some possible causes for the trends.

The necessity of precise signal location on scope faces is a problem of human operation which is critical in the technical methods of obtaining navigational information for the use of ships and commercial aircraft traveling in congested regions at night and in weather conditions adverse to good visibility.

History of the Problems

A signal can be located on a scope face by referring it to a system of scaling lines. The configurational and systematic errors of this procedure have already been reported by Ford (5, 6) and by Reed and Bartlett (9). The average random error for this method has been shown to be about 0.020 inch of scope distance for clear signals, and about 0.040 inch for typical unclear signals. The systematic errors are too great to be tolerated. Individual differences are relatively small.

Signals can also be located by a superimposed cursor which actuates a remotely controlled indicator. In this situation there are two human error problems: (a) how accurately the operator sets the cursor on the signal of "pip," and (b) how accurately he reads the indicator

dial. The first of these problems will be partially answered in the present article. With regard to the second, Chapanis (2, 3) and Long (7) have shown that a direct-reading counter is better than a scale with an indicator needle when reading the numerical value of the position of a signal which is already established as to location, but not for the purpose of moving an indicator to a specified numerical position.

Most of the war instruments used cursors with hair lines for centering the x- and the y-axes by separately controlled hand wheels. The use of a pantograph lever system for moving a pointer cursor with two degrees of muscular freedom was adopted by Reed and Bartlett (8) on PPI traffic survey scopes, and although they did not have a precision method for registering errors, they reported the centering procedure to be about twice as precise as the scale-reading method. This report is in the direction of our own findings on the sector scopes. In using cursors on changing target positions Ellson and Wheeler (4) reported the "range effect," which is a tendency of the operator to "overshoot" when the movement of a target position is less than expected, and to "undershoot" when the movement is greater than expected. Bartlett and Sweet (1) noted that when two operators were simultaneously tracking the same target their agreement in readings became poorer and poorer as the successive target positions moved more rapidly.

Apparatus

Figure 1 is a reproduction of an artificial signal (the oval area in the center) surrounded by eight station positions, actual size. In some of the experiments a photographic reproduction of a real radar "pip" was substituted for this artificial signal, at approximately the same size of signal area.

The cursor was a black dot the size of a No. 60 B and S drill inscribed on the under side of a transparent plastic plate. This cursor plate could be moved from any specified station position to the center of the signal by means of a pantograph system of levers such that experi-

* This research was executed under Contract No. W33-038-ac-22561 between the Institute of Research, Lehigh University, and the USAF Air Materiel Command, Aero Medical Laboratory, Wright-Patterson Air Force Base, Dayton, Ohio.

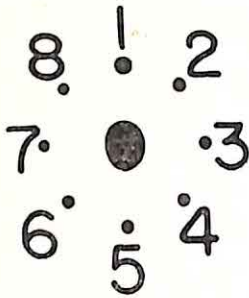


FIG. 1. Artificial signal, the oval area in the center. Numbered dots are the station positions from which the operator moves the point cursor (actual size).

mental ratios of knob movement to cursor movement could be established at values of 1:1, 2:1, 4:1 or 6:1.¹ The vertical error of

The target board was mounted thirty degrees from horizontal and illuminated with thirty foot candles of well diffused tungsten and fluorescent light.

Experimental Procedure

Method of Average Error. In this series the subject made his own settings of the cursor on the target. The order of the various conditions was determined by the Latin Square arrangement to equate fatigue and training effects from one condition to another.

The statistical treatment is determined by the fact that we discovered that all subjects tended to show "drifts" up and down for what they considered the center of the target area. A method of running averages and running standard deviations was adopted.

Table 1

Point Centering Experiments, Artificial Pip: Subject WEM
(Values expressed in decimal fractions of one inch)

Kind of Tracker	Ratio	Time Limit in Seconds	Stand. Dev. Total	Ave. Running S.D.	Max. Dev. Running Average	Drift Index
Pantograph	6/1	2	.0038	.0026	.0088	3.4
Pantograph	4/1	2	.0026	.0020	.0045	2.3
Pantograph	2/1	2	.0027	.0024	.0029	1.2
Fing. Cent.	1/1	8	.0028	.0020	.0065	3.3
Fing. Cent.	1/1	6	.0031	.0024	.0053	2.2
Fing. Cent.	1/1	4	.0031	.0025	.0049	2.0
Fing. Cent.	1/1	2	.0036	.0028	.0028	1.0

Discrimination Limen (Weber's Law), S.D. = .0010.
S.D., Reading from Scale A = .0166.

the centering was indicated by an attached rotating mirror which moved a light beam along a horizontal scale at a magnification of 100, and measured variations in settings to a precision of 0.001 inch.

The signal was covered by a solenoid-actuated shutter. The subject could open the shutter at his own discretion, but the shutter would close after an experimentally determined number of seconds allowed for centering of the target.

¹ For an expanded description of the apparatus, and additional tables and charts of results, see Ford, A., Rigler, D., and Dugan, G. E., Pantograph Radar Tracking: Point Centering Experiments, USAF Technical Report No. 5969, Air Materiel Command, Aero Medical Laboratory, Wright-Patterson Air Force Base, Dayton, Ohio.

Twenty-four trials were allowed for practice. Then about 180 settings were made in a series under a given experimental condition. The average and the standard deviation was computed for the first 24 readings. Then the first eight were dropped, and the next eight readings added, and the next average and standard deviation noted. This was done successively until the last 24 readings out of the 180 were exhausted. Table 1 and Figure 2 have been designed on this kind of data. Then the average and the standard deviation of the entire series was computed.

The drift phenomenon required quantification, so we established a "drift index" which was the maximum change in the running average divided by the average running standard deviation. If the operator's running average

shifted from a -0.003 to a $+0.003$ inch, and his average running standard deviation was 0.002 inch error, his drift index would be 3.0 . This example is fairly close to what actually happened. (See Figure 2.)

The Method of Constant Stimuli. Where the subject makes his own attempt to place the cursor in the center of a signal area one may consider three causes for the commission of an error: (a) the subject hasn't the motor skill to make a precise positioning of the cursor point, (b) the subject hasn't the fineness of visual perceptual discrimination to see where the center is, or (c) the subject lacks an attitude of working to precision and works only to the standard he thinks will satisfy the supervisor.

To get evidence on the second of the above factors, the experimenter set the cursor point in chance order above and below the actual center, and the subject reported "higher" or "lower" in a typical Weber-Fechner technique. The experimenter started with a plus-or-minus 0.004 , then 0.003 , and so on down to 0.001 inch differences. For this purpose the difference

limen was considered to be that value which the subject reported correctly by an amount equal to the standard deviation of the probability curve (68 per cent instead of the usual 50 per cent). The value of this limen, for WEM, is given at the bottom of Table 1. Such a value for each subject was assumed to yield a measure of the pure perceptual factor uncontaminated by the factor of motor skill.

Scale Reading Comparison. Using a set of photographic reproductions of artificial pips of the same dimension as that shown in Figure 1, superimposed on the multiple scaling system like that used in previous experiments which yielded the highest accuracy, each subject was required to estimate elevation of the signal above or below a line of zero reference, by the use of scaling lines only. Ninety-six readings were taken for each subject. The value is indicated under Table 1 after "S. D., readings from Scale A." This was done to show the difference in precision between scale-reading method and cursor-setting method, with individual differences held constant.

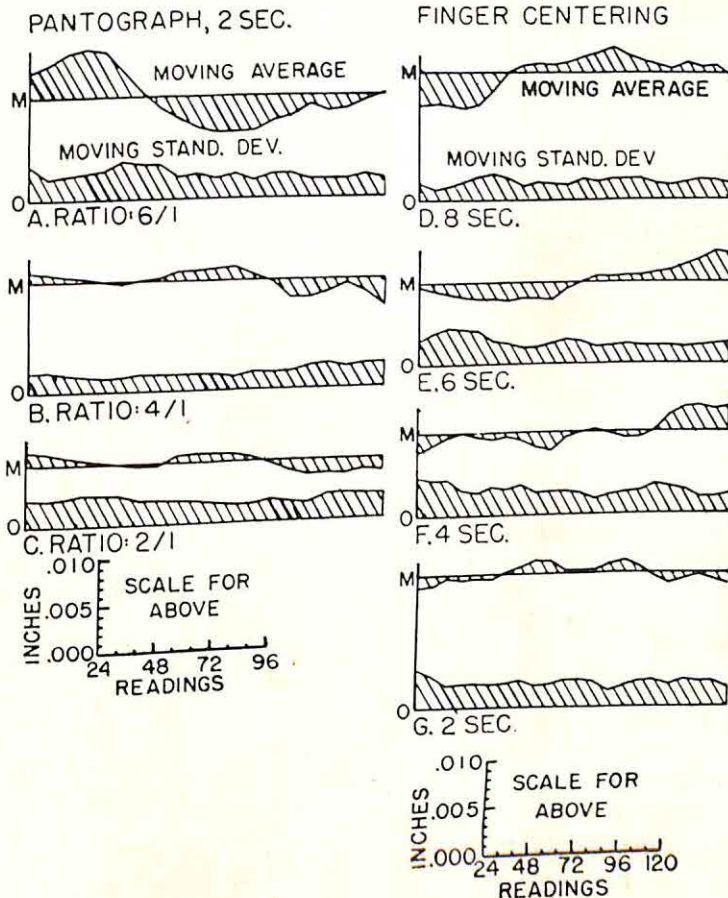


FIG. 2. Chart of error trends for Subject WEM on artificial signal.

Results

1. *Subjective Drift.* The tendency to wobble back and forth, above and below, the mean center, with the majority of readings too high for a while, then too low for a while, is pictured in Figure 1, for WEM. All seven subjects did this to some degree on both the artificial signals and the reproduction of the natural signal. The drift index was as much as *three* for every subject on one or the other of the runs. We are inclined to believe that this tendency to wobble around a center is characteristic of nearly any difference limen experiment by the method of average error, but has been heretofore hidden by the use of gross averages and gross measures of dispersion. About 0.003 inch wobble is the expectancy for this size of signal.

2. *Pantograph Reduction Ratio.* The best reduction ratio for hand movement in relation to cursor movement was close to 2:1. This is shown pictorially in Figure 2 for WEM, and in Table 1. The criteria for making this statement are: (a) least amount of drift, (b) smallest average running standard deviation of errors, and (c) smallest total standard deviation for errors. On the artificial signals, two out of the five subjects did almost equally well on the 4:1 lever ratio. On the reproductions of natural signals all three subjects were superior on the 2:1 ratio. Since the last situation is closer to the job requirements, the vote must go to the 2:1 ratio.

3. *Comparison to Scale Reading.* Using the total standard deviation (which includes variability due to drift) as the criterion for cursor centering by the 2:1 pantograph ratio, all five subjects showed a heavy superiority for pantograph centering as compared with the technique of scale reading. Expressing the standard deviation of cursor centering errors as percentage of the scale reading errors, we obtained the following values: 16, 46, 35, 44, 27. Over all, this means that the expectancy is that pantograph centering is about three times as accurate as reading the same signal on a superimposed scale (with much less mental effort). If a method of reducing drift could be found, the advantage would be still greater.

4. *Hand Positioning.* Allowing the subject

to rest the palm of his hand on the table while making a final centering resulted in a slightly superior score, statistically, but the difference was very small and of doubtful statistical significance. Subjects became rapidly antagonistic to the compulsory free arm method, without the palm rest, however, and reported that it was a strain to do it in this manner.

5. *Preceptive versus Motor Causes of Error.* The limen of perceptive discrimination, as indicated by the method of constant stimuli, is only a half to a third the value of the total standard deviation of cursor centering errors. This means that perceptive errors account for only a small part of the total error by the pantograph centering technique. We may probably make a large allowance for motor skills and precision attitude.

6. *Effect of Time Restriction.* A point cursor can be centered on a signal in a time as short as two seconds before any appreciable increase in error is generated. This was true of all eight subjects.

7. *Type of Cursor.* When the point cursor was surrounded by rings a tenth of an inch apart, like a "bull's eye," the drift tendency was reduced for all three subjects on the natural signal series. This suggests that the problem of drift can be solved by cursor design.

8. *Absolute Size of Errors.* The total standard deviation of pantograph centering errors for the 2:1 lever ratio (which includes the drift tendency) for the five subjects on the artificial signals was a plus-or-minus 0.0027, 0.0047, 0.0050, 0.0063, and 0.0026 inch, or an average expectancy of four thousandths of an inch for a signal of this size. The values for the centering of the reproductions of the natural pips were 0.0068, 0.0070, and 0.0050, or an average expectancy of six thousandths. Previous work by Ford (5, 6) has shown that natural signals are about twice as difficult to localize on a surface as artificial signals with sharp contours.

Summary

1. Centering signals on an area by means of a pantograph-controlled cursor is from two to four times as precise as the location of such signals by scale reading methods.

2. Using a point cursor, the subjects tend to exhibit systematic drift back and forth across the signal center.

3. A reduction ratio of two units of hand movement to one unit of cursor movement is near the optimum.

4. The use of a palm rest resulted in a slight superiority of scores in centering, and was greatly preferred by all subjects.

5. Centering in two dimensions can be done in a minimum of two seconds before reaching definite increases in error.

6. The limits of visual perception account for less than half of the range of errors. Skill and precision attitude may account for the remainder.

7. A point cursor surrounded by "bull's eye" rings reduces the tendency to drift around actual center.

8. With a signal 0.261 inch in length, the standard expectancy in centering is a plus-or-minus error of 4 thousandths of an inch for artificial signals, and six thousandths of an inch for reproductions of real radar pips. Individual differences are relatively large.

Received February 9, 1950.

References

1. Bartlett, N. R., and Sweet, A. L. *The agreement in range and bearing reports from two VF Radar Repeaters tracking the same target simultaneously with the B Scope*. Memorandum Report No.

166-I-12, Special Devices Center, U. S. Navy, 1 November 1946, restricted. Johns Hopkins University.

2. Chapanis, A. *Speed of reading target information from a direct reading counter type indicator versus conventional radar bearing and range dials*. Memorandum Report No. 166-I-3, Special Devices Center, U. S. Navy, 1 November 1946, restricted. Johns Hopkins University.
3. Chapanis, A. *The relative efficiency of a bearing counter and bearing dial for use with PPI Presentations*. Memorandum Report No. 166-I-26, Special Devices Center, U. S. Navy, 1 August 1947, restricted. Johns Hopkins University.
4. Ellson, D. G., and Wheeler, L. *The range effect*. Technical Report No. 5813, May 1949, Air Materiel Command, Wright-Patterson Air Force Base, Dayton, Ohio. University of Indiana.
5. Ford, A. Types of errors in location judgments on scaled surfaces. I. Errors of configuration. *J. appl. Psychol.*, 1949, **33**, 373-381.
6. Ford, A. Types of errors in location judgments on scaled surfaces. II. Random and systematic errors. *J. appl. Psychol.*, 1949, **33**, 382-394.
7. Long, G. E. *Speed and accuracy of readings as a function of design in the sensitive airspeed indicator*. Technical Report No. 5836, August 1949, Air Materiel Command, Wright-Patterson Air Force Base, Dayton, Ohio.
8. Reed, J. D., and Bartlett, N. R. *Comparison of manual and standard methods of target identification*. Memorandum Report No. 166-I-9, Special Devices Center, U. S. Navy, 1 February 1947, restricted. Johns Hopkins University.
9. Reed, J. D., and Bartlett, N. R. *The accuracy of range information on a PPI with respect to the distance of the target from the range rings*. Memorandum Report No. 166-I-29, Special Devices Center, U. S. Navy, 1 November 1947, restricted. Johns Hopkins University.

Influence of Friction in Making Settings on a Linear Scale *

William Leroy Jenkins, Louis O. Maas, and David Rigler

Lehigh University

In a previous study of the process of making settings on a linear scale by means of a control knob (1), a number of variables were investigated. The most significant of these turned out to be the *ratio* between pointer-movement and knob-turn. A ratio providing one or two inches of pointer-movement for one complete turn of the knob appeared to be optimal. Finer ratios wasted time and effort in traveling to the approximate location. Coarser ratios were poorly adapted to making the final adjustment.

In the apparatus employed for these studies there is a noticeable resistance with extremely coarse ratios. The question arose: Would the optimal ratio be changed by artificially equalizing the friction at all ratios? The present study is concerned with this question, and also with the more general problem of the influence of added friction at the optimal ratio.

Apparatus and Procedure

The apparatus and procedure were essentially the same as described in the previous study (1). The subject *matches* the position of a lighted insert in a black bakelite scale with a pointer controlled by a rotary knob. The permitted error-tolerance is determined by the width of the pointer in relation to the width of the lighted insert. In the present investigation, the permitted error tolerance was .007"; (insert .032"; pointer .025"), and the knob diameter $2\frac{3}{4}$ ".

By means of two chronoscopes, time is measured separately for travel to the approximate location and for making the final adjustment. Similarly, action potentials from the active forearm are accumulated and measured separately during travel and during final adjustment. Mean travel time is computed for two standard distances: 10 sixteenths and

50 sixteenths of an inch. Mean total time is then computed as mean travel time plus mean adjusting time. Similarly computations are made for action potentials, which are recorded in meter-scale readings having no intrinsic significance but comparable for the same subject.

In the present experiments, friction was added by means of a Prony brake on the shaft immediately behind the control knob. For the series on equalized friction, the braking was adjusted at each ratio so that a pull of 300 grams was required at the periphery of the $2\frac{3}{4}$ " knob—the same as for the coarsest ratio used. For the series on the influence of friction at the optimal ratio, sufficient braking was applied to require pulls of 100, 400, 700, 1,000, and 1,300 grams at the periphery of the knob.

Four subjects were used, all young men students. One of them (DMS) had served in several previous experiments. The other three were new.

Results—Equalized Friction

Table 1 shows mean total time and mean total potential at six different ratios without equalizing friction. These results are analogous to those reported in the earlier study (1). The optimal ratio is in the region of one or two inches of pointer movement for one complete turn of the knob. With a travel distance of 10 sixteenths of an inch, the ratio of 1.18 is optimal. When the travel distance is 50 sixteenths of an inch, the ratio of 2.42 is slightly better.

Table 2 gives mean total time and mean total potential for the same six ratios with friction artificially equalized. At 10 sixteenths travel distance, there is now little to choose between ratios 1.18 and 2.42. At 50 sixteenths travel distance, the ratio 2.42 becomes significantly better.

The two sets of data shown in Table 1 and Table 2 were taken at different times; so the absolute values cannot be directly compared.

* This research was executed under Contract No. W33-038-ac-22561 between the Institute of Research, Lehigh University, and the USAF Air Materiel Command, Aero Medical Laboratory, Wright-Patterson Air Force Base, Dayton, Ohio.

Table 1
Influence of Ratio on Time and Potential—Friction Not Equalized

Ratio	Mean Total Time							
	10 Sixteenths Travel				50 Sixteenths Travel			
	DMS	RBC	BRR	JMC	DMS	RBC	BRR	JMC
1.18	16.3	26.8	24.8	26.7	24.3	34.0	30.4	33.5
2.42	19.1*	27.5	25.3	32.6*	27.1*	31.5	29.3	36.6
4.08	19.2*	30.9*	26.4	35.2*	23.6	34.1	29.6	38.8*
6.28	19.5*	33.8*	28.8*	33.9*	23.5	35.8*	32.0	36.7*
9.70	23.8*	34.8*	31.6*	45.2*	26.6	37.2*	34.8*	48.4*
16.3	32.8*	48.8*	37.7*	52.5*	34.4*	50.4*	39.7*	56.1*

Ratio	Mean Total Potential							
	10 Sixteenths Travel				50 Sixteenths Travel			
	DMS	RBC	BRR	JMC	DMS	RBC	BRR	JMC
1.18	14.4	18.1	19.7	21.5	23.2	25.7*	28.5*	34.7*
2.42	17.1*	18.1	18.1	21.4	25.1	21.7	22.9	27.8
4.08	16.5*	21.0*	19.6	23.7	21.3	24.2*	23.6	29.7
6.28	18.1*	24.9*	23.6*	26.0*	22.1	27.7*	28.4*	30.0
9.70	19.7*	23.8*	27.0*	36.6*	23.3	26.2*	31.8*	41.0*
16.3	24.9*	38.1*	41.1*	45.1*	26.9*	40.9*	44.3*	49.5*

Tolerance: .007". Knob diam.: $2\frac{3}{4}$ ".

* Significant difference beyond 1% level from italicized figures.

Table 2
Influence of Ratio on Time and Potential Friction—Equalized at All Ratios

Ratio	Mean Total Time							
	10 Sixteenths Travel				50 Sixteenths Travel			
	DMS	RBC	BRR	JMC	DMS	RBC	BRR	JMC
1.18	26.8	27.8	20.2	40.4	41.2*	44.2*	28.6	55.2*
2.42	24.9	26.6	21.3	41.8	31.3	34.6	26.5	48.6
4.08	28.0	29.8	21.9	43.4	34.4	34.6	25.5	49.4
6.28	30.3*	30.8*	24.7*	47.5*	35.4*	35.2	27.7	52.3
9.70	31.9*	35.1*	27.4*	54.9*	36.5*	38.3	30.0*	59.3*
16.3	44.8*	38.6*	35.8*	70.9*	47.4*	41.8*	37.8*	73.3*

Ratio	Mean Total Potential							
	10 Sixteenths Travel				50 Sixteenths Travel			
	DMS	RBC	BRR	JMC	DMS	RBC	BRR	JMC
1.18	24.6	19.2	19.9	31.0	42.2*	36.4*	33.5*	46.0*
2.42	23.0	19.7	19.2	27.8	33.0	29.1	28.0	37.4
4.08	25.2	22.3*	19.7	32.1	32.4	27.9	25.7	39.3
6.28	28.3*	25.3*	21.9	33.9*	33.9	29.3	27.1	39.1
9.70	31.2*	33.8*	24.1*	41.1*	35.3	37.0*	27.7	46.3*
16.3	45.2*	29.9*	31.4*	59.3*	48.8*	33.1	34.6*	61.7*

Tolerance: .007". Knob diam.: $2\frac{3}{4}$ ".

* Significant difference beyond 1% level from italicized figures.

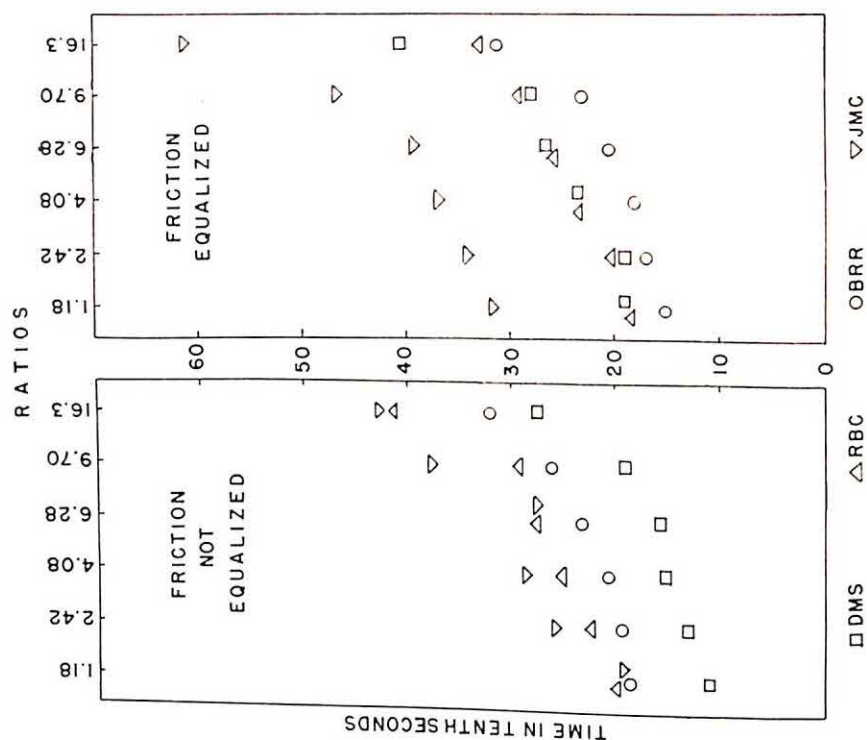


FIG. 2. Influence of ratio on adjust time.

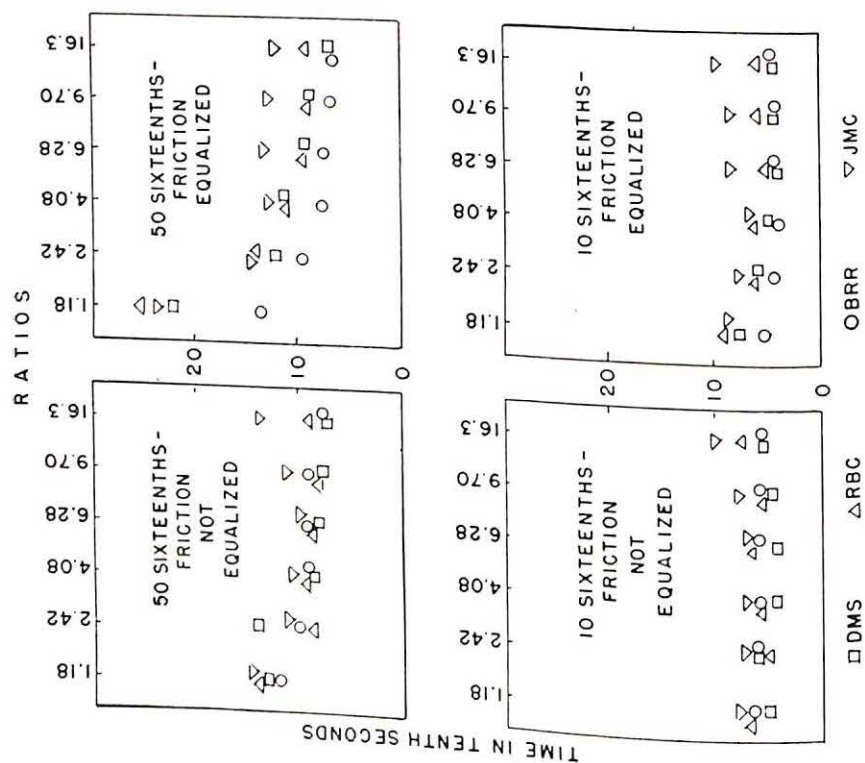


FIG. 1. Influence of ratio on travel time.

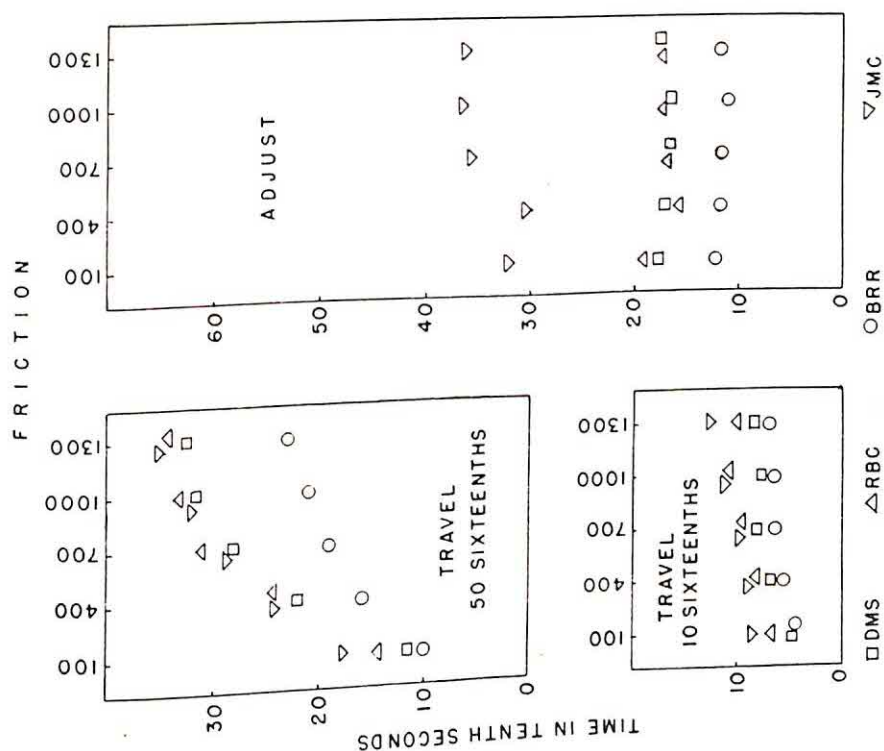


Fig. 3. Influence of friction on time.

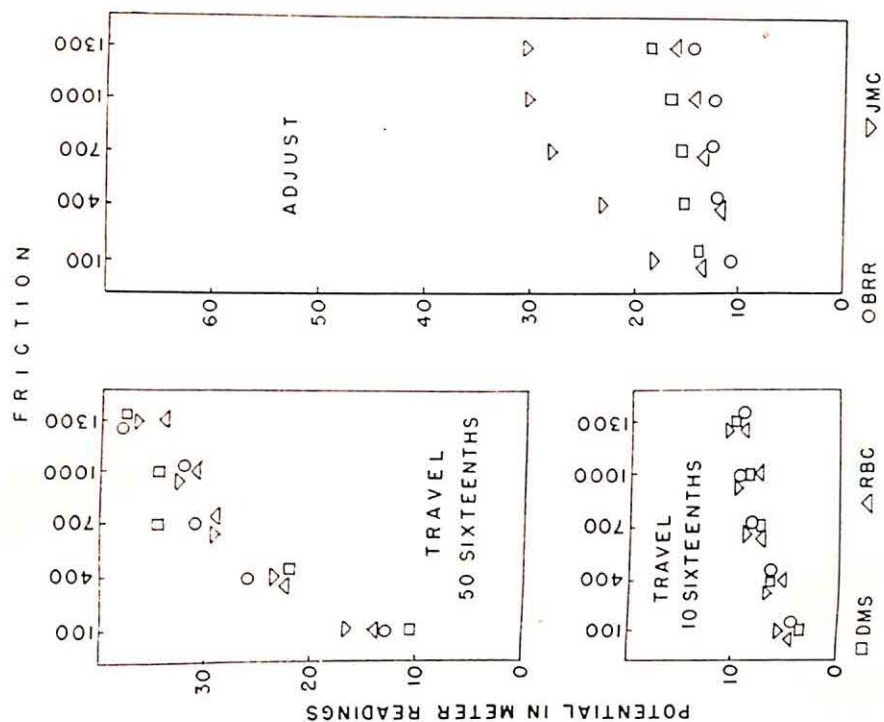


Fig. 4. Influence of friction on potential.

Table 3
Influence of Friction on Time and Potential at the Optimal Ratio (1.18)

Friction	Mean Total Time							
	10 Sixteenths Travel				50 Sixteenths Travel			
	DMS	RBC	BRR	JMC	DMS	RBC	BRR	JMC
100	22.4	26.1	16.6	40.3	29.2	33.7	22.2	49.5
400	23.8	23.9*	17.4	39.1	39.0*	40.3*	27.8*	54.3*
700	24.5	26.3	18.0	45.4*	44.5*	47.9*	30.8*	64.2*
1,000	23.7	28.0	17.6	47.6*	48.1*	50.2*	32.0*	68.4*
1,300	25.8*	27.6	18.8*	48.7*	51.0*	51.6*	34.8*	71.1*

Friction	Mean Total Potential							
	10 Sixteenths Travel				50 Sixteenths Travel			
	DMS	RBC	BRR	JMC	DMS	RBC	BRR	JMC
100	17.5	17.9	15.1	23.7	24.3	27.5	23.9	34.5
400	22.0*	17.0	18.9*	30.9*	37.2*	34.2*	38.1*	46.5*
700	23.0*	20.7*	21.4*	36.9*	49.8*	42.3*	43.8*	56.9*
1,000	25.1*	22.0*	21.9*	39.7*	55.1*	44.2*	44.3*	62.5*
1,300	28.5*	25.3*	23.5*	40.6*	56.1*	50.1*	52.3*	66.5*

Ratio: 1.18. Tolerance: .007". Knob diam.: $2\frac{3}{4}$ ".

* Significant difference beyond 1% level from italicized figures.

However, it is evident that equalizing the friction at all ratios does not change the region of the optimum at short travel distances and has only a moderate effect at longer travel distances.

Figures 1 and 2 show graphically why this is true. In Figure 1, travel time is compared with and without equalized friction. Only in the drop from 1.18 to 2.42 at 50 sixteenths travel is there any marked difference. In Figure 2, adjusting time is compared in a similar manner. With or without equalized friction, the ratio of 1.18 gives the shortest adjusting time. Any advantage of a higher ratio is, therefore, traceable to the slightly faster travel time over long distances. In most ordinary adjustments, the ratio of 1.18 remains optimal regardless of whether friction is artificially equalized.

Results—Added Friction at the Optimal Ratio

Normally in the apparatus a pull of approximately 100 grams at the periphery of the $2\frac{3}{4}$ " knob is required to move the pointer. By means of the Prony brake, friction was

added so that pulls of 400, 700, 1,000, and 1,300 grams were required.

Mean total time and mean total potential are shown in Table 3. At the shorter travel distance, even the highest level of friction increases the total time only slightly in three of the four subjects. The effect on total potential is more marked. At the longer travel distance, the slowing effect of friction is noticeable even at 400 grams and becomes progressively greater. Total potential increases in an analogous fashion.

Figures 3 and 4 show why these phenomena occur. In Figure 3 it will be noticed that added friction has its main effect on travel time, while adjusting time is scarcely affected at all. In fact, adjusting time with friction at the 400 gram level is slightly, though not significantly, shorter than at 100 grams. In Figure 4, travel potential shows a noticeable increase with added friction, but adjusting potential rises sharply in the case of only one of the four subjects.

In their reports, subjects do not complain about added friction at 400 grams but find

the higher level distinctly irksome even when no marked slowing of settings is evident.

Summary

Previously it had been shown that a ratio of one or two inches of pointer movement for one complete turn of a control knob was optimal for matching settings on a linear scale. In the present study, friction was artificially equalized at all ratios by means of a Prony brake. Even with equalized friction, the optimal ratio was unchanged for short travel distances. With large amounts of travel be-

fore setting, the optimal ratio was slightly increased. The influence of added friction at the optimal ratio was also studied. Added friction, even in excessive amounts, has no effect on adjusting time but does increase travel time and the action potentials from the active forearm.

Received February 9, 1950.

Reference

1. Jenkins, W. L., and Connor, M. B. Some design factors in making settings on a linear scale. *J. appl. Psychol.*, 1949, 33, 395-409.

The Galvanic Skin Response as a Test of Advertising Impact *

Edwin Golin and Samuel B. Lyerly

The University of North Carolina

In preparing an advertising campaign, the advertising agency is usually faced with the job of selecting the most effective layout for publication. The copy department submits several layouts concerning the product to be advertised and from this group, one arrangement with the greatest potential "impact" upon prospective consumers must be chosen. A misjudgment in this selective process may involve a large loss of capital, time, and energy. To avoid this waste, many techniques have been employed in an attempt to make the selection less haphazard. Some of the methods used are subjective judgments of large samples of population, experts' judgments, coupon returns, and so on. However, none of these methods has proved completely satisfactory because of either low validity or the relatively great expense and time factors involved. Experiences of this sort have placed the advertising industry on the alert for new methods of testing the impact of advertisements before the publication stage.

One possible way of approaching this problem is through the use of the galvanic skin response (GSR). Briefly, the GSR results from changes in skin resistance to an induced current. These changes of resistance may be read off a scale by deflections from a mirror galvanometer. It is assumed by many investigators that the GSR is a reflection of general autonomic nervous activity and that autonomic activity accompanies emotional states; hence GSR may be considered as one possible index of emotion. Applied to an advertisement, then, a layout of highly affective components for any particular observer should result in greater GSR deflections than a less affective one. The GSR method, if found valid, has the advantage of being a low-cost, time-saving technique.

With this in view, investigators have begun

* The authors wish to thank N. W. Ayer & Son, Inc., Philadelphia, for financial aid which made this study possible. Special thanks are due Mr. James M. Wallace, Vice-President, and Mrs. Margaret H. Rogers, Director, Copy Research Bureau, for their interest and cooperation.

experimenting with various types of GSR apparatus and advertising media. Among the limited amount of research in this area, the most prominent is the study by Eckstrand and Gilliland (1). These investigators found a significant agreement between GSR changes to a series of advertisements and the sales effectiveness of those advertisements. They concluded that the psychogalvanic method of testing advertising material could be a predictive technique if handled under properly controlled conditions.

It was because of this rapidly growing interest in GSR as a copy testing method that the authors in conjunction with an advertising corporation attempted to set up a preliminary factorial experiment that would indicate whether the GSR of a group of subjects would show significant differences when a series of advertisements concerning the same product was presented. This was to be a basic study and, as such, was not to involve other criteria such as subjective reports, sales tests, coupon returns, and so on. If the GSR did indicate significant differences among advertisements, then this could be used as the basis for further investigation and interpretation.

The Experiment

Advertising Material. Three products were chosen (nose tissue, ice cream, and air travel) which were considered within the range of interests of the subjects tested. These products were given fictitious brand names. Four advertisements were prepared for each product, varying the arrangement, or layout. Each ad consisted of a short slogan (three to six words), a picture, and the "brand" name. Layout A, for each product, featured prominently a picture of the article or service involved (for example, a large box of nose tissue or a picture of an airliner). Layout B included a picture of a single person using the advertised product (e.g., a boy eating an ice cream cone). Layout C portrayed two human figures, one male and one female, in poses which were

not directly related to the product. Layout D included pictures of miscellaneous content differing from Layouts A, B, and C in that the pictures were more elaborate and contained a much greater amount of detail. For all layouts, the slogans were appropriate to the picture and the product, and were devised in accordance with good advertising practice. The brand name was at the bottom of the advertisement in each case; but for each product, the positions of picture and slogan were alternated, the picture being at the top in half the ads, and the slogan at the top in the remaining half. The twelve ads were photographed in black and white, and 2×2 inch glass slides were prepared for each. These ads were supplied by the Copy Department of N. W. Ayer & Son, Inc.

Subjects. The subjects were sixty male undergraduate and graduate students ranging from eighteen to thirty-three years of age.

Apparatus. The GSR apparatus was similar to the alternating current circuit described by Grant (3). A Rubicon mirror galvanometer was used and the deflections accompanying changes in skin resistance were read off a scale graduated in centimeters.

Electrodes for the GSR apparatus were dry, polished silver discs attached to the palm

stabilization of GSR light deflections. Instructions were given to observe the advertisements as if the subject were glancing through a newspaper or magazine. A sample slide was presented followed by the twelve test slides.

The order of presentation was planned in accordance with the Latin square principle in such a way that over the group of 60 subjects each advertisement occupied a given ordinal position the same number of times. The twelve ads were arranged into five 12×12 Latin squares, and each of the 60 subjects was assigned at random to one of the rows of one square. The influence of contiguity upon responses to consecutive ads dealing with the same product or having the same layout was controlled by designing the Latin squares so that the product and layout effects were distributed systematically over the twelve ordinal positions. The first four ads in any row constituted one permutation of the four layouts, the second set of four was another permutation of the layouts, and the third set of four still another permutation. Similarly, the first and each succeeding set of three ads was a permutation of the three products. For example, the sequence presented to one subject was:

Order:	1	2	3	4	5	6	7	8	9	10	11	12
Advertisement:	3C	1B	2D	3A	2B	1D	2A	1C	3D	1A	2C	3B

and back of the right hand. Constant pressure was maintained by a clamp arrangement which, though comfortable, restricted excess hand movement.

The slides were projected on a daylight screen by a Keystone viewer operated on time exposure.

Procedure. The experiment was conducted in a dimly lighted room so situated that there was a minimum of distraction from the general activity of the building. The subject was comfortably seated approximately ten feet from the screen and so situated that the GSR apparatus and the viewer were not within his line of vision. The size of the image on the screen was approximately three feet square.

The electrodes were attached to the subject's hand and at least five minutes were allowed for adaptation of the subject and

Here the numbers 1, 2, 3 identify the three products, and the letters A, B, C, D, the four layouts; i.e., "3C" indicates that the first stimulus presented to this subject was the ad representing Product 3 in Layout C. The first group or block of four, comprising orders 1-4, has each of the four layouts represented once; and so do the two remaining blocks, 5-8 and 9-12. Orders 1-3 for this subject include each of the three products, as do orders 4-6, 7-9, and 10-12. The order of presentation was, of course, different for every subject, since 60 permutations of the basic sequence were used, each fulfilling the same conditions concerning the distribution of products and layouts.

The layouts were projected for four seconds, during which time the GSR deflections were recorded. There was an interval of approx-

imately thirty seconds between the presentations of successive stimuli.

Results

Table 1 contains the mean deflections in centimeters for each of the twelve advertisements individually, and the means for the product and layout groups. It will be noted that the mean deflections for the three products vary but little. The means for layouts, however, show considerably greater dispersion; the largest being associated with Layout C, which featured pictures of a young couple (a not unlikely result from a sample of college men), and the smallest with Layout A, the chief characteristic of which was a large picture of the product or service advertised.

As is often the case with data of this kind, distributions of the deflections were positively skewed, and there was a distinct tendency for the means and standard deviations among subjects to vary together. Applications of the usual statistical tests led to the rejection of the assumptions of normality and of homogeneity of variances and to the application of a logarithmic transformation to the data. The resulting distributions were more nearly normal and Bartlett's chi-square test of homogeneity (2, pp. 195-197) indicated that the null hypothesis could not be rejected at the 10% level. This degree of correspondence was considered adequate to warrant the application of analysis of variance.

Table 2 contains the results of the analysis. The residual mean square was used as the

Table 1

Mean Deflection (in Centimeters) for the Twelve Advertisements Classified by Product and by Layout*

Layouts	Products			Row Means
	1	2	3	
A	2.73	1.85	2.64	2.41
B	2.28	2.52	3.58	2.79
C	4.20	3.52	2.22	3.31
D	2.66	3.05	2.76	2.82
Column Means	2.97	2.74	2.80	Grand Mean 2.83

* N = 60 for each cell.

Table 2

Analysis of Variance of Logarithmically Transformed GSR Data

Source	Sums of Squares	d.f.	Mean Square	F
Products	.133	2	.066	.74
Layouts	.867	3	.289	3.25*
Subjects	18.880	59	.320	3.60**
Interactions:				
P × L	2.111	6	.352	3.96**
P × S	21.239	118	.180	2.02**
L × S	21.241	177	.120	1.35**
Residual	31.507	354	.089	
Total	95.978	719		

* Significant at 5% level.

** Significant at 1% level.

denominator in all F-tests. This residual mean square includes, in addition to "error" effects, any second-order interactions which may be present plus a component arising from response differences attributable to the order of presentation, or time sequence. The mean GSR plotted against order of presentation yields a negatively accelerated curve, dropping from a mean of 3.70 cm. on the first presentation to 2.05 cm. on the twelfth and indicating that some sort of adaption was in progress throughout the experiment. As a consequence of the experimental design, this circumstance operates primarily to inflate the residual mean square and has the effect of lowering all F's. It would be possible to make adjustments by taking account of the regression of GSR on order of presentation, but the conclusions drawn from the study would not be affected.

From Table 2 it will be noted that differences among subjects and differences among layouts are significant. Differences among products are, if anything, smaller than one would expect to find by chance, though not significantly so. All three of the first-order interactions are significant. The interaction of products and layouts is interpreted as indicating that, in the sample examined, the GSR response to an advertisement of a given product depends upon the way the advertisement is presented; i.e., that the most effective kind of layout for one product is not most

effective for another. The two interaction effects involving subjects indicate that response tendencies to different layout styles and to different products vary from individual to individual, depending upon differences in needs, experiences, preferences, or other individual characteristics.

The interpretations presented above should not be extended to apply to advertised products in general, to layouts in general, or to individuals in general. Neither the products nor the layout styles used are random samples of any clearly defined "populations" of products or layouts. The fact that the products were given fictitious brand names means, of course, that these findings cannot be generalized to advertisements of familiar brands toward which attitudes and response tendencies may be present. The sample of subjects (college men) is highly restricted, and yields no dependable predictions concerning responses to advertisements by people generally.

A factor whose effects cannot be precisely evaluated in this design is the brightness differences among the projected slides. These differences, which were considerable, were not anticipated, and no provision could be made for them in the factorial analysis. The brightness of each slide was measured with a Weston photometer and compared with the average GSR for that slide. The coefficient of correlation between brightness and average GSR is $-.06$, which indicates little or no relationship between GSR and brightness. Means of the brightness measures grouped according to product and layout do not demonstrate any significant differences. It is concluded that this variable did not unduly influence the experimental findings.

Summary and Conclusion

The desirability of a method for selecting the most effective advertisement from a group of preliminary layouts is recognized by advertisers, and recently some use has been made of the galvanic skin response (GSR) for this purpose. A factorial experiment was designed using the GSR in an attempt to differentiate the reaction-producing characteristics of layouts. Twelve ads were used, comprising four layouts for each of three products. The GSR responses of sixty male subjects to these advertisements revealed significant differences among layout styles and among subjects, but no significant differences among the three products. All first-order interactions of products, layouts, and subjects were significant.

The results of the study demonstrate that the GSR is sensitive to differences in layouts and that further research may lead to the development of practical techniques for testing the effectiveness of advertising copy before the costly publication stage is reached. The study also serves to illustrate the importance of careful planning and design of GSR research, including the use of proper experimental control and efficient statistical methods.

Received August 1, 1950.

Early publication.

References

1. Eckstrand, G., and Gilliland, A. R. The psychogalvanometric method for measuring the effectiveness of advertising. *J. appl. Psychol.*, 1948, **32**, 415-425.
2. Edwards, Allen L. *Experimental designs in psychological research*. New York: Rinehart and Co., 1950.
3. Grant, D. A. A convenient alternating current circuit for measuring GSR's. *Amer. J. Psychol.*, 1946, **59**, 149-151.

Book Reviews

Anastasi, A., and Foley, J. P. *Differential psychology* (Rev. ed.). New York: Macmillan, 1949. Pp. xv+894. \$5.00.

Anastasi's 1937 *Differential psychology* is now enlarged by four new chapters covering the basic concepts of psychological testing, the biological and psychological factors in simple behavioral development, and the effects of schooling on intelligence. Other material, accrued since 1937, on trait organization and socio-economic differences has been added to the earlier discussion and is here reorganized into two substantially new chapters. Eighteen other chapters which conform closely to their former headings and sub-headings, have been re-written to include researches of more recent date. Outstanding inter-disciplinary contributions are certain studies from genetics, anthropology, and sociology.

Sectional additions, scattered throughout the book, comprise: the measurement of special aptitudes, conditions which affect the shape of the distribution curve, the nature of heredity, the nature of environment, the heredity-environment relationship, popular misconceptions regarding heredity and environment, structural and functional characteristics, the concept of unlearned behavior, the study of practice as an approach to the heredity-environment problem, typical findings on the improvement of mental test performance in age, the constancy of the I.Q., the study of family pedigrees, the search for components of physique and temperament, constitutional type as a social stereotype, profile analysis, what intelligence tests measure, a cultural concept of intelligence, evaluation of group differences, the comparative achievement of different races, and the cultural frame of reference in behavior, developmental stages, language, and human nature.

The book is divided to orient the reader historically and methodologically, to present an analysis of individual differences, and then to discuss group differences. Studies from sociology and anthropology are integrated with research in psychology to give a larger perspective on common problems.

This second edition reflects a more studied consideration of content and aptness of expression, and shows growth in unraveling theoretical complexities. The obscured operation of heredity and environment in group differences is analyzed in a more mature fashion, attention being focused on newer insights from recent methodological refinements in the construction and application of psychological tests, from the constantly accruing longitudinal studies, and from the interest expressed in the fundamental nature of intelligence and personality through analyses of their specific components.

The new edition affords a useful compendium for professional workers and students who cannot themselves read, digest, and coordinate original research publications. The main service of this book, however, is in the training it gives in being alert for "common pitfalls and sources of error in interpretation of obtained results," with the authors repeatedly puncturing conclusions drawn from insufficient data, unreliable tests, poorly designed research experimentation, ineffective controls, inadequate statistical treatment, or just plain immaturity of psychological understanding. This second edition is well documented, well illustrated, well arranged, and well written.

Gladys C. Schwesinger

California Youth Authority

Calhoon, Richard P. *Problems in personnel administration*. New York: Harper and Brothers, 1949. Pp. xii+540. \$4.00.

The emphasis in this book is on "the understanding of personnel problems: the reasons therefor, with methods of attack and solution so that the feel of actual problems can be more real to the student of personnel administration whether he is in the field or in the classroom." With this objective in mind, Calhoon presents a critical evaluation of present developments in personnel policy and programs.

Through the twenty chapters in the book, Calhoon treats the various kinds of problems faced by personnel administrators, classified in the usual way around selection, training, wage and salary administration, etc. For each phase of a personnel program, Calhoon at-

tempts to show its place in the total program, to point out common errors or weaknesses of present practice, and to make suggestions for a sound program. Practical problems arising in actual administration are stressed throughout. So, for example, in discussing "executive training," Calhoon judges the practice of a one-to-two year rotating program for college graduates as a "dubious procedure at best," points out difficulties such as maintaining motivation and getting the trainee accepted by line supervisors and employees, and then presents the case for having general training follow rather than precede a definite job assignment. Literally hundreds of personnel problems are discussed in this fashion. The book thus becomes a field manual for personnel administrators and a guide for those in training for personnel administration.

As such, its value depends upon the validity of the analyses, criticisms, and suggestions made. Unfortunately, there is inadequate research behind many of the opinions and beliefs upon which personnel policies and practices are now decided. To supplement his own opinions and experience, however, Calhoon reports the results of a questionnaire survey of nearly 600 personnel administrators on nearly all of the major personnel problems. In addition, the NICB surveys and AMA publications are frequently referred to. It can be assumed that, right or wrong, here is a distillation of the best opinion in the field, at least as viewed and reported by one person.

To facilitate its use in personnel training situations, either in school or in industry, Calhoon has ended each chapter with practical discussion problems, projects, and role-playing situations. These study aids are well chosen and, if used properly, will vitalize the use of the book as a text. Its greatest weakness, however, is the lack of reference material since there is no attempt to survey the supporting evidence in the literature and there are no lists of suggested readings. Particularly neglected are references to psychological literature; only one psychological journal article and only a few texts in industrial psychology are referred to and those but briefly.

Albert S. Thompson

Teachers College,
Columbia University

Stuit, Dewey B., Dickson, Gwendolen S., Jordan, Thomas F., and Schloerb, Lester. *Predicting success in professional schools*. Washington, D. C.: American Council on Education, 1949. Pp. v+187. \$3.00.

This book summarizes present knowledge gained from experience and research in predicting success in professional schools. It is edited under the sponsorship of the Committee on Student Personnel Work of the American Council on Education and is based on a series of technical bulletins prepared by the Veterans Administration for use by vocational advisers. Professional fields included are engineering, law, medicine, dentistry, music, agriculture, teaching, and nursing.

The already comprehensive compilation of information in the Technical Bulletins has been extended in its usefulness by a very good editing job. A valuable introductory chapter has been added on problems and techniques of prediction. This contains such a good discussion of the problems of prediction, and limitations of research, that it would merit reading and re-reading by all practitioners. Mention is made of the absence of predictive knowledge relative to such factors as interest, motivation, and personality, but this is not, in the writer's opinion, strongly enough emphasized, since the burden of effort is first to establish the importance of what quantitative findings exist. There is, however, a good discussion of the dearth of investigation relative to the influence of environmental factors in the prediction of professional success. The authors wisely caution continuously, not only in this chapter, but also in the concluding statement for each profession, against applying generalizations for prediction to individual cases.

The general discussion of tests as predictive indices omits raising the question of whether tests measure what they purport to measure. It is also somewhat disappointing that although there is emphasis on the restriction of the use of clinical instruments (i.e., personality tests) to professional workers with adequate background in Clinical Psychology, there is not equal caution expressed regarding the interpretation and clinical use of any testing in the counseling process.

For each profession helpful general informa-

tion concerning national requirements precedes the presentation of research findings. The way in which this material is organized for presentation is particularly good. For example, the effect of length of pre-training, native vs. transfer training, etc., are considered separately, followed by an evaluation of each kind of test and combination of indices. A summary and comment on implications for counseling completes each section. The book is in highly readable form, an excellent job having been done in sifting out and simplifying the important findings. There is a useful table of essential qualifications for each field of professional training, with a suggested battery of tests. These test recommendations are open to question, especially recommendations given in the personality area. In addition, no mention is made of the fact that, in generalizing regarding the lack of predictive value of interest inventories, such as the Strong, only a single key has been used, when it is generally considered that patterning or constellation is more diagnostically significant.

Despite these few omissions, this book is an invaluable compilation for any high school or college counselor. It would seem also that in carefully selected cases it might profitably be made available to counselees who are considering one or another of the professional fields covered. In any case, filling a major gap, it is an imperative addition to the working materials of the professional counselor.

(Mrs.) Barbara A. Kirk

Counseling Center,
University of California,
Berkeley, California

Schramm, Wilbur, Editor. *Mass communications*. Urbana: University of Illinois Press, 1949. Pp. xi, 552. \$4.50.

Mass communications is a collection of readings. It claims to be nothing beyond that, and the sub-title clearly states that it is a book of readings. Consequently, there is no reason to expect that prospective purchasers will be misled into expecting anything else.

There are several indications that this collection of readings fills a real need, rather than just being another book. Experience with the symposium, *Communications in modern society*, of the Institute of Communications Research

of the University of Illinois suggested the need for something that would be broader in scope. Furthermore, there were a number of more recent publications that deserved consideration; so the Director of the Institute selected and edited this collection of readings. Since all of the selections have been used in the Institute, this publication rests on a background of actual experience.

In a field as broad as this one, anyone who attempts to select and edit a limited number of readings obviously is confronted by several very difficult problems. His success in solving these problems largely determines the value of the publication.

Providing a satisfactory degree of continuity is one of these problems. Schramm's solution consists mainly of grouping the readings and giving a short introduction for each section. The section on the development of mass communications includes readings on the history of the newspaper, motion pictures, radio and international communications. The control and support of mass communications are considered in a second section. The third section consists of five readings on various aspects of the communications process from the standpoints of general semantics, psychology, and anthropology. The three remaining sections cover content, audiences, and effects.

The appendix presents several pages of a wide variety of facts and figures.

Judged by conventional standards used to evaluate most publications, this collection of readings would have to be rated as spotty in coverage and below par in continuity. However, it should not be judged by such standards. The gaps are the result of the characteristics of this field and are beyond the control of any editor. The raw material for an integrated treatment of the subject just does not exist.

Perhaps an even more serious problem is the headache of deciding how much space to give each paper. Complete reproduction of each article and chapter selected would be possible only by limiting coverage to a point at which the objective would not be accomplished.

The longest reading is the summary of Borden's *The economic effects of advertising*, which is 41 pages. Second in length is Llewellyn White's *Ragtime to riches* (28 pages), which

is one of the two readings by the same writer on the development of American radio. Only two other readings are longer than 20 pages. Twenty-three readings run ten to twenty pages. Fifteen readings, not counting the section introductions, are shorter than ten pages.

Inclusion of the shortest readings raises a question. For example, the shortest reading is one page of definitions of the devices of propaganda from *The fine art of propaganda*. Another reading consists of three tables showing changes in heroes, settings, and themes in large-magazine fiction (from Johns-Heine and Gerth). The summary of Kenneth Baker's study of radio programming takes up only two pages, and so does the presentation of the Flesch readability formula.

At first glance these short readings appear to contribute to the impression of inadequate continuity. In addition, such treatment of a subject might be misleading: However, including them does increase the coverage of the volume as a whole. Since the book probably will be used most often in class situations that permit further explanation and since individual users outside of classes probably will consult the original sources, the reviewer agrees with Schramm's decision to include them.

When complete reproduction is not practical, one is faced with a problem of what to omit. In several such cases, Schramm reports results and conclusions at the expense of description of the methods. Although this may tend to cause some concern among technicians, it is the only practical approach which is in line with his objectives. Obviously, this book was not intended for research technicians in a strict sense. Many of the readings are what tough-minded research people probably would consider "essays," although some of them are relatively quantitative. This is as it should be. Any other approach would present a distorted picture of the broad field of mass communications.

The usefulness of this book depends on the need that any particular person might have for a collection of readings on this subject. It provides convenient material for class purposes, and it at least minimizes the amount of digging necessary to get an introductory overall picture of this field. There is ample evidence of both good judgment and experience in the selection and editing. Any deficiencies result from the

nature of the available material and the weaknesses inherent in any publication which is based upon the selection and editing of previously published manuscripts.

Alfred C. Welch

*Knox Reeves Advertising, Inc.,
Minneapolis*

Massing, Paul. *Rehearsal for destruction*. New York: Harper and Brothers, 1949. Pp. 341. \$4.00.

The book is well-written, carefully documented, and scholarly in style, but is strictly a "who did what" (with dates) historical analysis. The absence of any systematic discussion of the group or individual psychological determinants and factors in German anti-Semitism in the period covered (1871-1914) leaves little of interest except for the most zealous student of the subject. The broader purposes of the study may well have required this painstaking chronicling approach, but the result is a volume of only slight interest to most psychologists.

Harrison G. Gough

*University of California,
Berkeley, California*

McCord, Carey P., and Witheridge, William N. *Odors: Physiology and control*. New York: McGraw-Hill, 1949. Pp. 405. \$6.50.

In these days of ever-increasing specialization it may appear as temerity to attempt covering in one volume most of what is known about odors—from the anatomy of the olfactory system and the physiology of olfaction to the legal aspects of odor nuisances; from the theoretical considerations of the mode of action of the chemical stimuli to presentation of practical procedures for odor cancellation and counteraction; from classification of the odors to the use of odors as warning agents in fuel gases and as a method of broadcasting alarm in mining emergencies; from discussion of the rather unsuccessful efforts at clarification of the relationship between chemical structure and the quality of odors to the control of some industrial odors by altering the chemistry of the manufacturing processes; from odor of the human body, in health and disease, to household odors and odors of water and foods. The

authors were concerned, primarily, with the offensive odors and the engineering of odor control. The applied psychologist will read with some interest the chapters on industrial odors and the offensive trades. He might be consulted, on rare occasions, about making an odor survey; he will find in the present volume valuable practical suggestions.

Because the area covered by the book is so large, many topics are treated rather sketchily. This appears to be particularly true of matters psychological. Thus, in connection with the question of pleasantness and unpleasantness of odors, mentioned only very briefly, the authors feel that "Such matters so invade the realm of emotions and habituations that further delineation must be delegated to the psychologists" (p. 28). The short section on "mental perturbation" resulting from offensive odors (p. 71) is little more than a sample of "literary psychology." The experimental psychologist will note the new electrochemical theory of odor stimulation (p. 22), based on Linus Pauling's ideas about molecular architecture and biological reactions in general and odor sensations in particular, but he will be aware of its highly speculative nature.

There is a large and moderately useful bibliography (124 pp., some 3,500 references), arranged by chapters. The authors stated that several hundred of the items were cited from secondhand sources. It would have been useful to identify this material by some simple symbol and to separate it from references actually seen.

University of Minnesota

Josef Brožek

Bellows, Roger M. and Rush, Carl H., Jr. *Workbook in personnel methods*. Dubuque, Iowa: Wm. C. Brown Co., 1949. Pp. 102. \$2.10.

This workbook is designed for use in an introductory course in personnel methods to accompany Bellows' "Psychology of Personnel in Business and Industry." It attempts to "enable the personnel methods trainee actually to participate in personnel methods problems." It emphasizes the objective approach to problems and gives practice in the handling of quantitative data and in quantitative approaches to human relations problems.

In line with this orientation, the authors have prepared exercises which require the analysis and interpretation of data such as might be obtained in an actual situation and which illustrate typical personnel techniques. So, for example, an exercise on the reliability of criteria deals with inter-rater agreement, one on recruitment requires computing average test scores of applicants from three recruitment sources, one on selection involves an item analysis of application blank items, one on training requires plotting learning curves on two groups of industrial trainees, etc. Practice is given in the use of Taylor-Russell tables, an interview guide, an item discrimination nomograph, norms, scattergrams, a merit rating form, a sociogram, the Flesch readability formula. Altogether 22 exercises related to 14 of the 20 chapters in the text are provided.

In general, the exercises are well worked out; the instructions are clear, the data are well presented, the interpretation questions are thought-provoking. The common danger of a workbook's becoming merely a study guide to the text has been avoided. It is not the "case-study" type of textbook supplement but one which is designed to give the student an exposure to some of the simpler tools used in the scientific approach to personnel problems.

Albert S. Thompson

Teachers College,
Columbia University

Pressey, Sidney L. *Educational acceleration: appraisals and basic problems*. Columbus, Ohio, Ohio State University: Bureau of Educational Research, 1949. Bureau of Educational Research Monographs, No. 31. Pp. xiv + 154. \$2.50, paper. \$3.00, cloth.

This book, which should be of interest to all who are concerned with policies in all levels of educational institutions, presents the results of a research project conducted at Ohio State University. According to the author, "The effort was made broadly to investigate a problem, experimentally to try various means for dealing with it, and systematically to appraise the results." In this, Pressey has produced a stimulating and valuable piece of work.

Until the last war resulted in pressure for educational acceleration, discussion of it and attempts to try it were largely characterized

by *a priori* thinking and little data. The changes brought about by the war, however, provided an opportunity to study the effects of acceleration on large numbers of students. The author did this extensively and considered the problems carefully.

Pressey has shifted the emphasis to the question of what justification there might be for continuing to prolong adolescent status and delay mature social functioning and productivity. He has found little or none. The work is organized around three basic concepts which lend clarity and direction to the studies: maturity, prime (defined as peak biological potentiality), and individual differences.

After chapters dealing with these basic concepts and with the opinions and studies of other educators and psychologists, the rest of the book seems to fall into three main sections: one part deals with the trend toward delayed completion of training; another presents the data bearing on the results of acceleration; the third is concerned with alternative methods of acceleration.

In the latter two sections, which are the more important, Pressey shows that, contrary to popular notions, accelerated students tend not only to be superior in academic achievement (even when matched with non-accelerates with equal aptitude), but they take part in as many social activities, and, after graduation, make equally good or better job and life adjustments. Included in this section is a study demonstrating the need for *guidance* in selecting students for acceleration.

The major alternative method of acceleration (in addition to summer attendance and heavier course loads) suggested is credit by examination. This is shown to have no adverse effects on later scholarship. In addition, data are presented, taken from questionnaires and interviews, which show that in many cases, the accelerates feel that the present most common methods place them under undue strain. The conclusion drawn is that before even guided acceleration can be effective, educational institutions will have to revise their rules and mores to avoid making the accelerates feel out of place or too pressed.

The most serious criticism of this book lies in the area of the presentation of the data. The conclusions are based on differences be-

tween percentages of various groups meeting certain criteria. Yet nowhere is there presented a test of the significance of these differences. Pressey says, for example, "The differences are so small as to be of questionable significance, each taken by itself. Taken together, they show a consistency . . ." While such consistency may argue for the acceptance of the results, still professional readers will probably have as hard a time as the reviewer in understanding why so eminently able a person as Pressey should have left this crucial step in the analysis undone or unreported.

Pressey concludes the book by reemphasizing the need for guidance in any flexible program; he then makes eight specific recommendations which, he feels, might serve as guideposts to people willing to come to terms with the three basic concepts which he presented at the beginning.

Whatever the shortcomings of the analysis, Pressey has done a masterful piece of work in clearing the ground so that the real issues can be clearly seen; other investigators may well be able to explore the problems further and more adequately because of his work.

John W. Gustad

Vanderbilt University

Weston, H. C. *Sight, light and efficiency*. London: H. K. Lewis and Co., Ltd., 1949. Pp. xiv + 308. 42s.

This English treatise, based largely upon the results of recent research sponsored by the Medical Research Council and the Department of Scientific and Industrial Research in Great Britain, is concerned mainly with illumination in relation to industrial and occupational efficiency. After preliminary attention to the sense of sight and causes and symptoms of eye-strain, consideration is given to occupational demands upon sight, the facilitation of visual tasks, lighting and visual efficiency, incentive luminance and color, testing vision, and protection of sight at work. Recommended levels of illumination are listed in the appendix. There is an abundance of appropriate figures, tables and illustrations.

There are many contributions in this book which may be cited with approval. Among these are the following: (1) The industrial psychologist, the illuminating engineer, and

others interested in visual efficiency will find material of practical value. (2) To a large degree the cited experimental data provide a sound basis for the conclusions presented. (3) One should know what acuity is needed for comfortable work rather than acuity just sufficient to permit the job to be done. (4) The extent to which contrast and size are interdependent factors affecting the severity of the visual task is clearly explained. (5) It is stated (page 170) that "however high illumination is made, objects which are not very nearly alike either in apparent size or contrast, or in both these characteristics, never become equally easy to see." These findings should help to correct the views of those American writers who claim that the effects of small size and poor contrast can be completely compensated for by increases in illumination. (6) Methods of illuminating work objects are effectively described. (7) The psychological feeling of well-being that comes with adequate illumination combined with desirable combinations of color and brightness of painting receives a desirable emphasis. (8) For the most part, the schedule of recommended values of illumination are adequate (not excessive) in terms of experimental findings.

Careful evaluation will raise questions on certain specific details. A few of these criticisms follow: (1) During recent years it has often seemed practical to compute the illumination level needed for a specific visual task. A job analysis reveals size of detail and brightness contrast. Then by reference to laboratory studies of visual acuity in relation to size of detail, contrast and illumination, the level of illumination for the job is computed. Weston has worked out nomograms to facilitate such determinations. Although this procedure is helpful at the present stage of our knowledge, it is a dubious practice. It is questionable whether it is valid to make a direct transfer from visual acuity data to a relatively complex seeing situation where integrated visual patterns are involved. Furthermore, recommendations which rest on estimates from threshold measures have not in most cases been validated for the actual visual work activities. When this has been done, the computed illumination values have been

found to be greater than needed. (2) When the author cites the need of 500 foot candles or more for lace-making because workers prefer to work out in the sun rather than in poorly illuminated interiors (page 153), his inference is unwarranted. There is no other choice, such as a well lighted interior, available. (3) It is possible that other factors than differences in illumination (page 162) affect the shape of the work curves in winter as compared with summer. (4) The tentative inference (page 142) that white print on black background is more readable than black on white is not supported by experimental findings.

Industrial psychologists and illumination engineers should welcome this well organized and well written book on the relation between illumination and visual efficiency. Careful reading should reveal how extravagant some of the illumination intensities recommended in this country are.

Miles A. Tinker

University of Minnesota

Ward, Roswell. *Out-of-school vocational guidance*. New York: Harper and Bros. 1949. Pp. xiv + 155. \$2.50.

Admittedly, the literature of vocational guidance dealing specifically with the "out-of-school" population is limited. The author draws upon extended experience as a vocational counselor and as an organizer of out-of-school vocational guidance services in many states, in an attempt to fill the gap.

"Vocational guidance" is held to be a fact-finding, coordinating, planning and referral service. The "vocational counselor" who is the chief figure in this service, does more than counseling (the "planning" part of vocational guidance). His effectiveness is determined most by his skill as a fact-finder and his ability to coordinate and make use of information and resources.

The narrowing of the traditional definition of "vocational guidance" shows its value both in the organization and in the operation of a vocational guidance service. Many practical ideas are presented for carrying on an organized community service to meet the continuing need for vocational guidance.

The author does not hesitate to present rather dogmatically certain ideas which his

own experience has convinced him are valuable. Among these, the following are typical:

The identification of the "non-selective" worker permits economies through the use of less intensive forms of counseling.

Adherence to a classification of "vocational problems" according to symptoms is consistent with the defined function of the vocational counselor as a maker of plans for vocational adjustment and as an "action-centered" rather than "applicant-centered" fact-finder and referral agent. The categories of this classification (each preceded by the word "Vocational") are: Immaturity, Confusion, Insecurity, Misdirection, Fixation, Conflict.

A sharp distinction between "interest" and "motivation" has implications for the fact-finding activities of vocational counseling. An eighteen-item classification of vocational motivations is presented.

An insistence that vocational counseling must be tied basically to vocational information and vocational economics rather than to psychology, has implications for the selection and training of vocational counselors. Prejudice is revealed here which is carried to the point that the training of vocational counselors would be taken out of schools of education and departments of psychology. The author likens the contrast between vocational guidance as an adjunct of education and as an out-of-school service, to the difference between fire prevention and fire fighting.

A criticism of "mirror counseling" is a further revelation of the same prejudice. Consistent with the fear of "mirror counseling" is a distrust of a psychologist's ability to become a vocational counselor. The psychologist should be only a resource person available to the "applicant" (not "client") by referral from the counselor.

One of the limitations (other than the fairly obvious effects of the revealed prejudices) of the volume lies in the heavy dependence upon the "applicant" either in the interview or by filled-in forms, for the data about himself essential to his vocational planning. The usual categories of data are mentioned but techniques for securing and evaluating them are left largely to the judgment, attained chiefly through experience, of the vocational counselor.

It is the opinion of the reviewer that when the author's prejudices are fully discounted, the book stands as a practical challenge to many traditional practices. Also, the book should be thought-provoking, rather than offensive, to psychologists and educators concerned with carrying on counseling services and with the preparation of counselors.

Fred M. Fowler

*Department of Public Instruction,
The State of Utah*

Allport, G. W. *The individual and his religion.*
New York: The Macmillan Company, 1950.
Pp. x + 147. \$2.50.

Almost independently of its merits, this book is important in three respects. First, and I think of greatest significance, is the fact that an American psychologist of first rank and theoretical interests has addressed himself to this topic. As Allport points out, contemporary psychologists avoid the subject of religion about as much as James and his contemporaries avoided any detailed treatment of sex. (Sexuality, by the way, gets strikingly little attention in Allport's discussion of the organic "origins of the religious quest," pages 9-12.) The state of mind of modern Western man, including many of the intellectuals, is such that even the most atheistic psychologist, if interested in personality, ought to concern himself with the religious impulses in somewhat more detail than to dismiss them airily as "sublimations," "failure of nerve," and the like. That a writer of Allport's stature has done so may stimulate others, who do not share his scientific views about personality, to do the same. Secondly, this book will be useful to counselors as bibliotherapy for clients experiencing religious conflict, especially in the case of students whose conflicts arise via contact with psychology and allied sciences. Thirdly—and here I part company with Allport—the book is a challenge to the determined and wholehearted secularist in psychology, because it is quite obviously *not* a disinterested psychological causal analysis but includes a not-too-subtle bit of religious propaganda, the author's avowal (p. vii) to the contrary notwithstanding. I do not know the facts about Allport's own religious convictions, if any; but unless I am mistaken he has a

religious urge two sigmas above the American psychologists' mean. Let me cite a few random examples having this flavor. "The universe is simply incomprehensible. Fragments of it may be fairly well understood, but certainly not the design of the whole" (p. 18). "The fact . . . is that scientific thought is known . . . to be able to cover only part of the ground. It goes a long way but not far enough" (p. 20). ". . . even in this age of technology and social disintegration when skepticism hangs heavily upon the horizon. . ." (p. 39).

Certain purely "psychological" theses are presented without the qualifications one would like to see in a book that will get into the hands of non-psychologists. On page 63, we find that ". . . in only slight degree, if at all, is this energy drawn from the reservoir of organic drives"—treating Allport's important and respectable theory of autonomy as an established principle of psychological science. We are assured (p. 65) that if a strong religious sentiment fades away, ". . . we can be certain that religion was never a central feature of the personality" (Allport's form of Calvin's doctrine of perseverance, perhaps?).

The book is marred by a few minor mistakes, such as Allport's canonizing of Thomas à Kempis (p. 124), and the lugging in of that dreadful old chestnut, so often refuted by careful logical analysis, about determinism's incompatibility with praise, exhortation, and effort (p. 115). The presentation of the positivist critique under "referential doubting" (pp. 117-121), especially the end of the paragraph at the top of page 119, is incredibly superficial. Positivists do not "condemn" poetry, for it does not claim to assert facts. "Legal discourse" is primarily a metalinguistic examination of "what entails what," and is, so far as I know, admitted by most positivists. The lumping together of "poetic, artistic, legal" discourse as meaningless to positivists shows a carelessness which in a less respected writer would almost suggest a lack of intellectual integrity.

Finally, the book fails to satisfy because of a deep-seated methodological mistake. It assumes that a really well-rounded psychological discussion of religion *can* be given without any concern with the truth-character of

religious beliefs. By this device, Allport tries to have his cake and eat it too; he grinds the religious axe and at the same time avoids the necessity of defending any particular religious thesis. But certain of his "psychological" assertions can simply not be made unless the prior question of validity has been settled. For example, he several times rejects the common characterization of religion as "wish-fulfilling." But he is at great pains to exhibit its role in satisfying needs. Now, if the objects of (propositional!) religious belief do not exist, or if the evidence for their existence is slight, those who have such beliefs, and have them on the psychological bases Allport describes, are surely engaging in wish-fulfilling, whether Allport likes the phrase or not! The constant use of evaluative terms and remarks shows that Allport is not carrying out a *purely* causal analysis, but he never comes to grips, except in a very sketchy and dilettante manner, with the cognitive issues *about* religion which are involved. The book suffers here from that subjectivism, that making man the measure of all things (I am sure Allport would deny this!) which Bertrand Russell has described as the "characteristic madness of our times." For this reason, the philosophically sophisticated clinician could use the book therapeutically only by permitting himself an essentially cynical view of the religiously conflicted student—"it will help him, because he won't see through it."

Paul E. Meehl
University of Minnesota

Tinkelman, Sherman. *Difficulty prediction of test items*. New York: Bureau of Publications, Teachers College, Columbia University, 1947. Pp. 55. \$1.85.

Using a civil service examination for patrolman consisting of 100 5-option multiple-choice items covering arithmetic reasoning, civic information, logical inference, reading comprehension, judgment, vocabulary and other topics, the author had thirty judges rate the difficulty of each item. This rated difficulty (per cent of patrolman candidates expected to answer correctly) was compared with the actual per cent of patrolman candidates passing. The judges "seemed to constitute a professionally competent group of exami-

ners" with from three to thirty years of experience with the examining agency. The validity of the judges' estimates of difficulty (correlation between estimated and actual per cent correct) ranged from .23 to .77 with a median of .53.

The validity of the pooled judgment of the thirty judges was .76. Nine graduate students in an advanced test construction course, (presumably without three to thirty years experience) repeated the estimates for an entirely different test. For this trained, but inexperienced group (on a different test, to be sure), the validities ranged from .17 to .72 and the validity of the pooled judgment (9 judges) was .81, higher than the value of .76 obtained from 30 experienced personnel examiners.

Analysis of variance techniques were used to identify factors affecting the predictability of item difficulty. Predictability was unrelated to item content—difficulty of vocabulary items was no more accurately predicted than was that of arithmetical reasoning and the differences among content categories could have arisen by chance. Judges were relatively consistent in their ability (or inability) to estimate or guess the difficulty of the items as measured by their performances on odd and even items.

Even within the limits of "validity" represented by a median coefficient of .53, there were large constant errors of over- or underestimation. For only nine of the thirty judges were these constant errors statistically insignificant. Since the constant errors were in both directions, the pooled judgment was substantially accurate.

The judges tended to overestimate the percentage of passes on difficult items and to

underestimate it for easy items. The judgments thus exhibited the usual regression toward the mean. There were no significant differences in predictability between "judgment" and "information" questions or between items with high and those with low item-test coefficients (erroneously termed validity coefficients).

The problem is introduced in a civil service setting—and a rather conservatively conceived one. The necessity for basing the selection of items on subjectively judged estimates of difficulty is mentioned on page 1, but *nowhere is the need for estimates of validity—subjective or statistical—alluded to*. The impossibility of experimental tryout of items is ruled out because of the need for secrecy although progressive merit systems have found ways of tryout without violating security.

The tone of the conclusions and implications seems unwarrantedly optimistic. "The results of this study indicate that judges are probably able to estimate relative item difficulty accurately." This conclusion does not seem warranted for a median validity correlation of .53. Moreover, although the "best" three judges have a combined validity of .85, the author fails to note that this is a "back-validated" coefficient, comparable to that computed for a test on the item analysis sample after the best items have been selected by item analysis.

Although Tinkelman's book does shed light on the process of judging difficulty, it will not convince the careful examiner that estimated difficulty is an adequate substitute for empirical data.

Charles I. Mosier

*Adjutant General's Office,
Washington, D. C.*

New Books, Monographs, and Pamphlets

Books, monographs, and pamphlets for listing and possible review should be sent to Donald G. Paterson, Editor, Department of Psychology, University of Minnesota, Minneapolis 14, Minnesota

- Projective psychology.* Lawrence E. Abt and Leopold Bellak, Editors. New York: Alfred A. Knopf, Inc., 1950. Pp. 485. \$6.00.
- Converting a veterans guidance center.* George D. Barahal. California: Stanford University Press, 1950. Pp. 99. \$1.50.
- Beginning experimental psychology.* S. Howard Bartley. New York: McGraw-Hill Book Co., Inc., 1950. Pp. 483. \$4.00.
- Problem-solving processes of college students.* Benjamin S. Bloom and Lois J. Broder. Chicago: University of Chicago Press, 1950. Pp. 109. \$2.75.
- Readings in modern methods of counseling.* A. H. Brayfield, Editor. New York: Appleton-Century-Crofts, Inc., 1950. Pp. 526. \$5.00.
- Sex questions and answers.* Fred Brown and Rudolf T. Kempton. New York: McGraw-Hill Book Co., Inc., 1950. Pp. 256. \$2.95.
- Scientific research: its administration and organization.* George P. Bush and Lowell H. Hattery, Editors. Washington, D. C.: American University Press, 1950. Pp. 190. \$3.25.
- Personality, a systematic, theoretical and factual study.* Raymond B. Cattell. New York: McGraw-Hill Book Co., Inc., 1950. Pp. 689. \$5.50.
- The application of measurement to health and physical education.* Second edition. H. Harrison Clarke. New York: Prentice-Hall, Inc., 1950. Pp. 455.
- Recent experiments in psychology.* Second edition. Leland W. Crafts, Theodore C. Schneirla, Elsa E. Robinson, and Ralph W. Gilbert. New York: McGraw-Hill Book Co., Inc., 1950. Pp. 503. \$3.50.
- Some theory of sampling.* William E. Deming. New York: John Wiley and Sons, Inc., 1950. Pp. 602. \$9.00.
- Psychology.* Second edition. Dockeray and Lane. New York: Prentice-Hall, Inc., 1950. Pp. 576. \$3.75.
- America begins.* Richard M. Dorson, Editor. New York: Pantheon Books, Inc., 1950. Pp. 438. \$4.50.
- The counseling interview.* Clifford E. Erickson. New York: Prentice-Hall, Inc., 1950. Pp. 174. \$1.75.
- Communicating ideas to the public.* Stephen E. Fitzgerald. New York: Funk and Wagnalls Co., 1950. Pp. 292. \$3.50.
- The principles of scientific research.* Paul Freedman. Washington, D. C.: Public Affairs Press, 1950. Pp. 222. \$3.25.
- The psychology of dictatorship.* G. M. Gilbert. New York: Ronald Press Co., 1950. Pp. 327. \$4.00.
- Outside readings in psychology.* E. L. Hartley, H. G. Birch, and R. E. Hartley, Editors. New York: Thomas Y. Crowell Co., 1950. Pp. 880. \$2.75.
- Clinical applications of suggestion and hypnosis.* William T. Heron. Springfield: Charles C Thomas, Publisher, 1950. Pp. 125. \$3.00.
- Textbook of abnormal psychology.* Revised edition. Carney Landis and Marjorie M. Bolles. New York: The Macmillan Co., 1950. Pp. 634. \$5.00.
- Psychiatry for social workers.* Second edition. Lawson G. Lowrey. New York: Columbia University Press, 1950. Pp. 385. \$4.50.
- In my mind's eye.* Frederick Marion. New York: E. P. Dutton and Co., Inc., 1950. Pp. 315. \$3.75.
- Carbon dioxide therapy, a neurophysiological treatment of nervous disorders.* L. J. Meduna. Springfield: Charles C Thomas, Publisher, 1950. Pp. 248. \$5.00.
- Experiments in social process.* James Grier Miller, Editor. New York: McGraw-Hill Book Co., Inc., 1950. Pp. 201. \$3.00.
- Physiological psychology.* Second edition. Clifford T. Morgan and Eliot Stellar. New York: McGraw-Hill Book Co., Inc., 1950. Pp. 572. \$5.00.
- Psychology and art of the blind.* G. Revesz. New York: Longmans, Green and Co., Inc., 1950. Pp. 338. \$8.50.
- The development of a needs and problems inventory for high-school youth.* Benjamin Shimberg. Lafayette, Ind.: Studies in Higher Education LXXII, The Division of Educational Reference, Purdue University, 1950. Pp. 78.
- Delinquency and human nature.* D. H. Stott. Fife, Scotland: Carnegie United Kingdom Trust, Dunfermline, 1950. Pp. 460. \$1.00.
- Improving management communication.* General management series No. 145. New York: American Management Association, 1950. Pp. 26. \$75.
- Practical operating problems in personnel administration.* Personnel series No. 129. New York: American Management Association, 1949. Pp. 28. \$75.
- Industrial applications of medicine and psychiatry.* Personnel series No. 130. New York: American Management Association, 1949. Pp. 31. \$75.
- German aviation medicine, World War II.* Vol. I and II. Prepared under auspices of the Surgeon General, U.S.A.F. Washington, D. C.: Superintendent of Documents, U. S. Government Printing Office, 1950. Pp. 1302. \$8.50, set.
- Wages, employment and personnel problems in a changing economy.* Production series No. 187. New York: American Management Association, 1949. Pp. 71. \$1.25.